ACCV

北京科技大学
University of Science and Technology Beijing

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Tutorial in 17th ACCV 2024.12.09

From Machine-Machine Comparison to Human-Machine Comparison: Adapting Visual Turing Test in Visual Object Tracking

Prof. Xin Zhao (赵鑫)

- Professor in University of Science and Technology Beijing (USTB)
- https://www.xinzhaoai.com/
- xinzhao@ustb.edu.cn

Dr. Shiyu Hu (胡世宇)

- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

Scan to download this tutorial PPT

ACCV 2024

Organizers

Prof. Xin Zhao (赵鑫)

- **Professor in University of Science and Technology Beijing (USTB)**
- **https://www.xinzhaoai.com/**
- **xinzhao@ustb.edu.cn**

Prof. Xin Zhao received his PhD degree from the University of Science and Technology of China (USTC) in 2013. His research interests include video analysis and performance evaluation, especially for object tracking tasks. He has published the international journal and conference papers, such as the IJCV, IEEE TPAMI, IEEE TIP, IEEE TCSVT, CVPR, ICCV, NeurIPS, AAAI, IJCAI. Recently, he has mainly conducted research on human-computer vision evaluation. He has built several widely-used computer vision benchmarks (i.e., GOT-10k, VideoCube, SOTVerse, Biodrone, etc.) with online evaluation platforms. He has regularly served as program committee member or peer reviewer for the following conferences and journals: CVPR, ICCV, ECCV, ICML, NeurIPS, ICLR, IJCV, IEEE TPAMI, IEEE TIP, IEEE TMM, etc.

ACCV 2024

Organizers

Dr. Shiyu Hu (胡世宇)

- **Research Fellow in Nanyang Technological University (NTU)**
- **https://huuuuusy.github.io/**
- **shiyu.hu@ntu.edu.sg**

Dr. Shiyu Hu received her PhD degree from the University of Chinese Academy of Sciences in Jan. 2024. She has authored or coauthored more than 20 research papers in the areas of computer vision and pattern recognition at international journals and conferences, including TPAMI, IJCV, NeurIPS, etc. Her research interests include computer vision, visual object tracking, and visual intelligence evaluation.

ACCV 2024

CONTENTS

- Introduction
- Part 1. Visual Object Tracking Task
- Part 2. Experimental Environments
- Part 3. Algorithms and Traditional Machine-Machine Comparisons

Time Break (about 20 minutes)

- Part 4. Human Visual Abilities and Visual Turing Test
- Trends and Future Directions

Scan to download this tutorial PPT

ACCV 2024

CONTENTS

- **Introduction**
- Part 1. Visual Object Tracking Task
- Part 2. Experimental Environments
- Part 3. Algorithms and Traditional Machine-Machine Comparisons
- Part 4. Human Visual Abilities and Visual Turing Test
- Trends and Future Directions

Scan to download this tutorial PPT

ACCV 2024

Introduction

- What is Visual Object Tracking?

Visual information: 83%

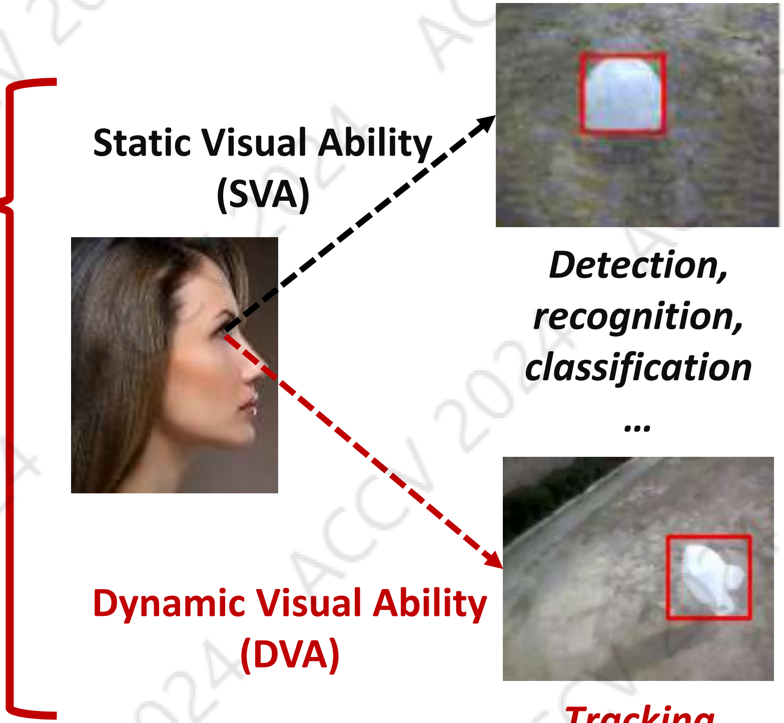
Auditory information: 11%

Olfactory information: 3.5%

Taste information: 1%

Tactile information: 1.5%

Humans are "visual animals"



Static Visual Ability (SVA)

Detection, recognition, classification …

Dynamic Visual Ability (DVA)

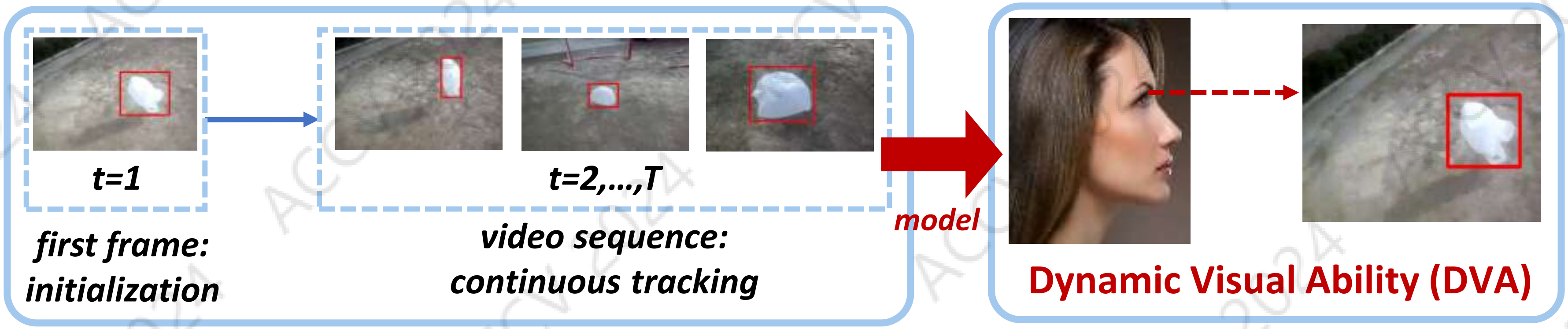
Tracking

Visual object tracking (single object tracking) is a basic function of the human dynamic visual system.

ACCV 2024

Introduction

- **What is Visual Object Tracking?**


t=1
first frame: initialization
t=2,…,T
video sequence: continuous tracking
model
Dynamic Visual Ability (DVA)

➢ **Definition:** Provides only the initial position of a moving object, and continuously locates it in a video sequence.

➢ **Characteristics:**

- **Sequential decision:** locating the target with the help of previous frames
- **Category agnostic:** without any assumption about the target category (open-set setting)
- **Instance-level prediction:** need to distinguish the target from others (including objects in the same category)

ACCV 2024

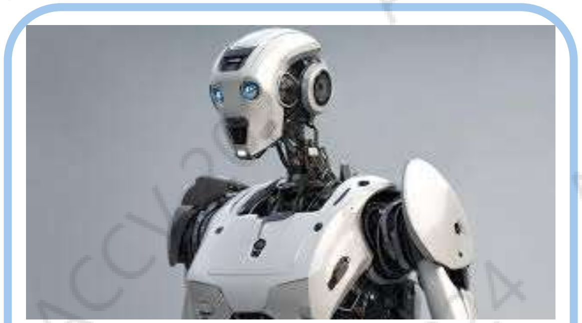
Introduction

- ## Why is VOT Important?

➢ **Real-world demands**: More **intelligent and robust** visual tracking systems are needed to adapt to complex real-world environments.

Autonomous driving: Tracking vehicles or pedestrians to ensure road safety.

Video surveillance: Real-time tracking of suspicious targets in security systems.



Robot vision: Robots track objects through vision systems to interact with the environment.

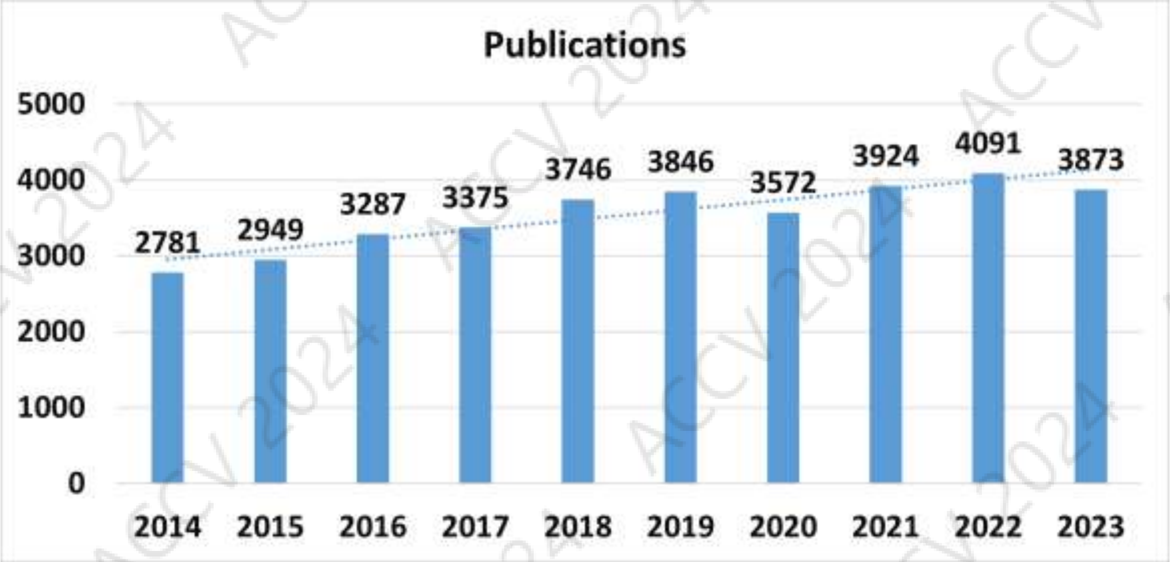
ACCV 2024

Introduction

- **Why is VOT Important?**

➢ **Academic hotspots:**

- The graph shows a consistent growth in publications related to "object tracking" over the past decade. Despite minor fluctuations, the overall trend is upward, indicating **sustained interest and ongoing research** in the field.
- This growth reflects the **increasing attention** to visual object tracking, driven by advancements in deep learning and AI technologies.



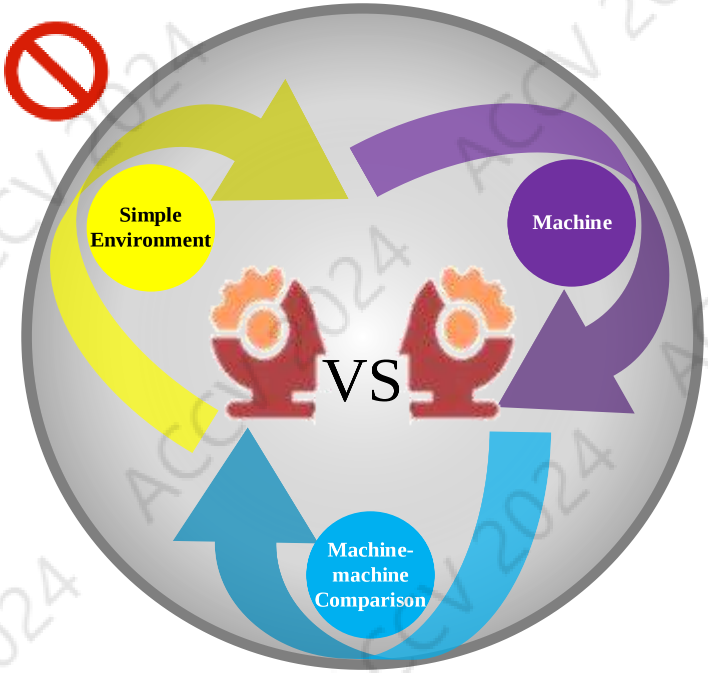
Increasing number of publications in object tracking from 2014 to 2023. (Data from Web of Science, allintitle: object tracking.)

ACCV 2024

Importance of Evaluation Techniques

- **Lagging Evaluation Techniques**

➢ **Inadequate Current Evaluation Standards:** Most evaluations **focus on performance but ignore intelligence**. The evaluation is restricted to **machine-to-machine comparisons within simple environments.**



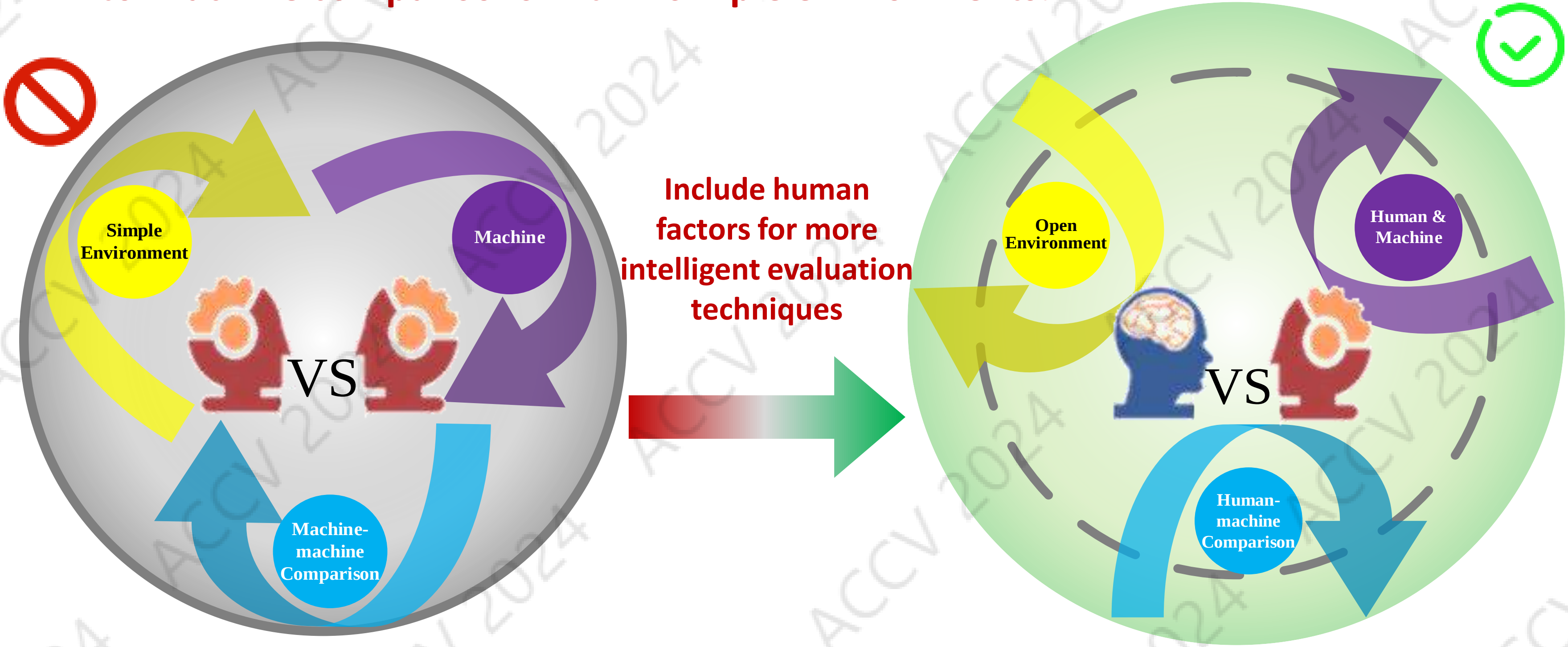
- **Environment:** Current evaluations often use **simple and controlled** environments, which fail to represent the complexities of real-world scenarios.
- **Executor:** The focus is mainly on the **machine's performance**, with little consideration for human capabilities.
- **Evaluation:** Most systems rely on **machine-to-machine comparisons**, which do not fully reflect human-level intelligence or decision-making processes.

ACCV 2024

Importance of Evaluation Techniques

- ## Lagging Evaluation Techniques

➢ **Inadequate Current Evaluation Standards:** Most evaluations **focus on performance but ignore intelligence**. The evaluation is restricted to **machine-to-machine comparisons within simple environments.**


Simple Environment
Machine
VS
Machine-machine Comparison
Open Environment
Human & Machine
VS
Human-machine Comparison

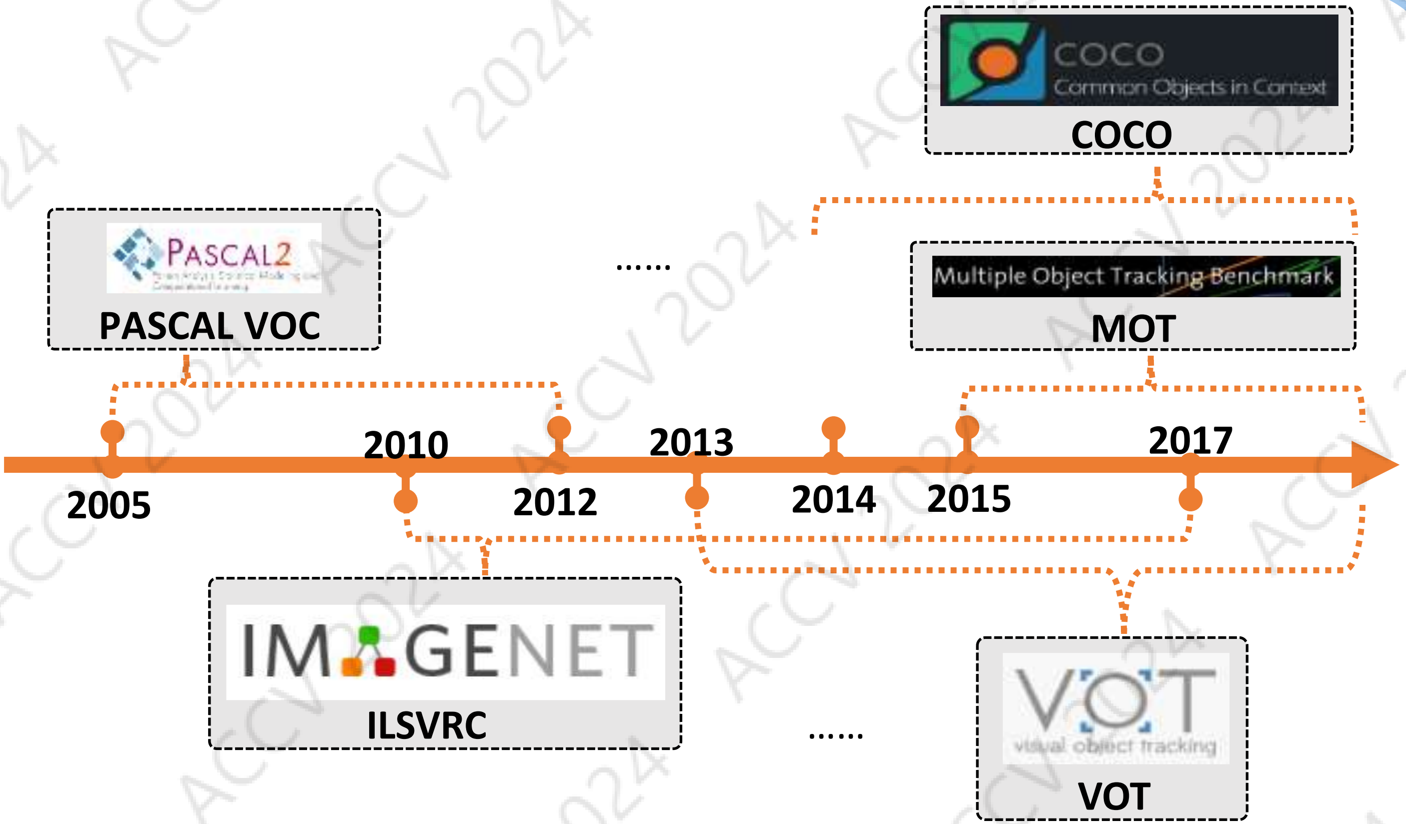
Include human factors for more intelligent evaluation techniques

- The environment is more **open** and reflects real-world complexities.
- There is **comparison between humans and machines**, making the evaluation process more holistic.

ACCV 2024

Importance of Evaluation Techniques

- ## Importance of Visual Intelligence Evaluation
  - **Performance Bottlenecks of Algorithms:** Evaluation techniques can **reveal weaknesses** in algorithms across different scenarios and **provide feedback for design optimization**.


COCO
Common Objects in Context
COCO
PASCAL2
PASCAL VOC
......
Multiple Object Tracking Benchmark
MOT
2005
2010
2012
2013
2014
2015
2017
IMAGENET
ILSVRC
......
VOT
visual object tracking
VOT

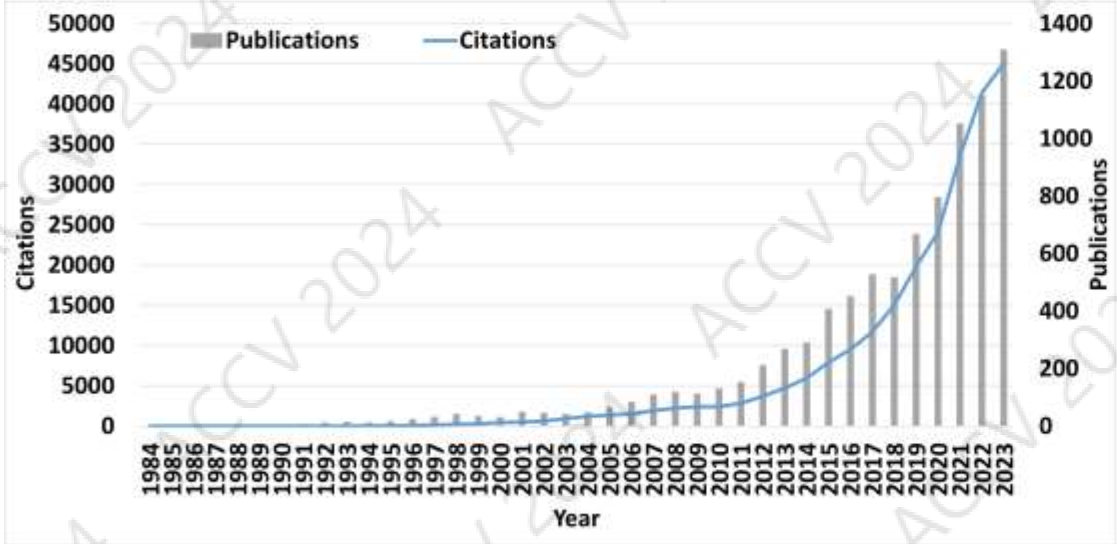
Famous competitions that have advanced computer vision

ACCV 2024

Importance of Evaluation Techniques

Importance of Visual Intelligence Evaluation

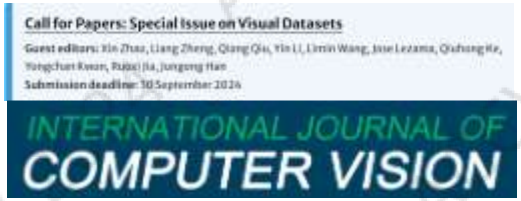
- **Future Outlook**:The advancement of AI and visual intelligence depends on advanced evaluation techniques. Only through **scientific evaluation** can we **ensure that advancements** in AI and visual intelligence continue at the current pace. Evaluation not only drives innovation but also helps ensure that algorithms meet the complex demands of real-world applications.



Increasing number of publications and citations in computer vision evaluation from 1984 to 2023. (Data from Web of Science, allintitle: computer vision evaluation.)

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS datasets & benchmarks track from 2021



Call for Papers: Special Issue on Visual Datasets

Guest editors: Xin Zhao, Liang Zheng, Qiang Qiu, Yin Li, Limin Wang, Jose Lezama, Qiuhong Ke, Yongchan Kwon, Ruoxi Jia, Jungong Han

Submission deadline: 30 September 2024

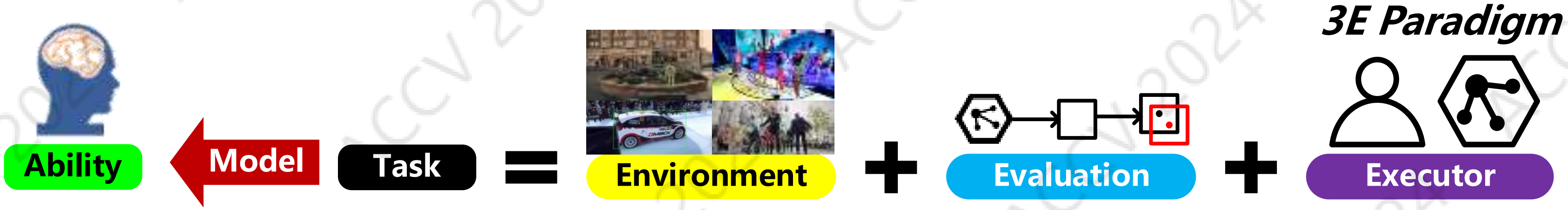
INTERNATIONAL JOURNAL OF COMPUTER VISION

Special Issue on IJCV

ACCV 2024

Structure and Goals of the Tutorial

- **Structure**
  - **Introduction (this section)**
  - **Part 1. Task Definitions and Challenges**: A deep dive into the foundations and challenges of object tracking.
  - **Part 2. Categorization of Evaluation Environments**: Examination of the environments used for evaluating tracking algorithms.
  - **Part 3. Algorithms and Traditional Machine-Machine Comparisons**: Analysis of existing algorithms and their performance comparisons.
  - **Part 4. Human Visual Abilities and Visual Turing Test**: Discussion on human-machine comparison frameworks and the introduction of the Visual Turing Test concept.
  - **Trends and Future Directions**



3E Paradigm

Ability ⟵ **Model** **Task** = **Environment** + **Evaluation** + **Executor**

S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," *International Journal of Computer Vision* (**IJCV**), 2024.

ACCV 2024

Structure and Goals of the Tutorial

- **Goals**
  - Help participants understand visual intelligence **evaluation techniques** in visual object tracking.
  - Discuss the **strengths and weaknesses** of existing evaluation mechanisms and offer directions for improvement.
  - **Inspire future research** in visual tracking.

Expected Takeaways:

Through this tutorial, participants will gain a **comprehensive understanding** of visual intelligence evaluation techniques.

ACCV 2024

CONTENTS

- Introduction
- **Part 1. Visual Object Tracking Task**
  - **Task Definition**
  - **Task Challenge**
- Part 2. Experimental Environments
- Part 3. Algorithms and Traditional Machine-Machine Comparisons
- Part 4. Human Visual Abilities and Visual Turing Test
- Trends and Future Directions

Scan to download this tutorial PPT

ACCV 2024

Task Definition

Explore the fundamental definitions that shape how we approach human visual tracking ability.

ACCV 2024

Task: Short-term Tracking (STT)

Visual/Single Object Tracking (VOT/SOT)

- **Characteristics: Sequential decision, category agnostic, instance-level prediction.**

model

Human Visual Tracking

hidden constraints

Characteristics of STT (based on VOT challenge):

- **Single-target**
- **Model-free**
- **Causal relationship**
- **Short-term**
- **Single-camera**

extra task constraints

A STT demo (≈30s, in single camera)

Definition: Short-term tracking refers to **continuously tracking** a single object within a short sequence, where the **target remains visible** in every frame of the video. It assumes no significant interruptions, occlusions, or camera changes.

Early research simplified the task, which is far away from human visual tracking ability.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

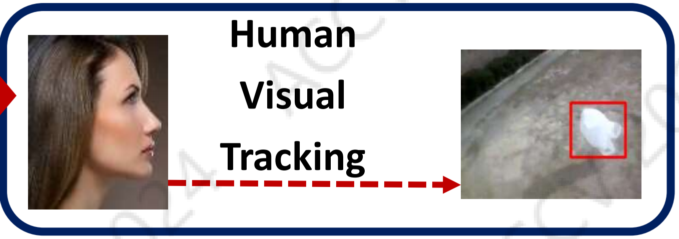
ACCV 2024

Task: Long-term Tracking (LTT)

Visual/Single Object Tracking (VOT/SOT)

- **Characteristics: Sequential decision, category agnostic, instance-level prediction.**

model



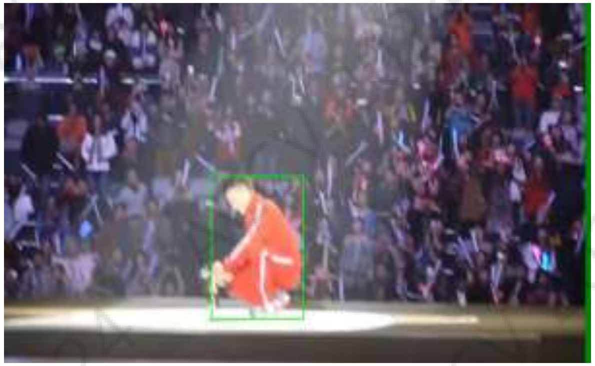
Human Visual Tracking

still hidden constraint (single-camera)

Characteristics of LTT (cancel short-term constraint):

- **Single-target**
- **Model-free**
- **Causal relationship**
- **Single-camera**

extra task constraint



A LTT demo (≈2mins, in single camera)

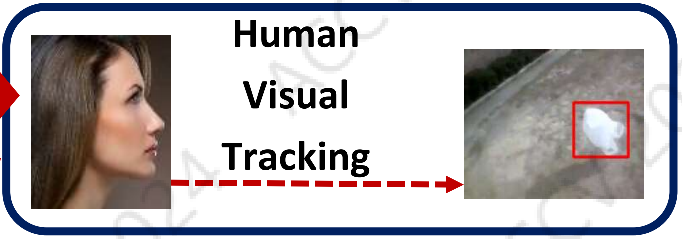
Definition: Long-term tracking expands single object tracking to longer time sequences, allowing for **temporary disappearance** of the target (due to occlusion or leaving the frame) and **requiring re-detection** when the target reappears. This is in contrast to short-term tracking, which assumes the target is always present in the frame.

Task: Global Instance Tracking (GIT)

Visual/Single Object Tracking (VOT/SOT)

- **Characteristics: Sequential decision, category agnostic, instance-level prediction.**

model



Alignment

Characteristics of GIT (cancel all extra constraints):

- Single-target
- Model-free
- Causal relationship

No.000676
Shotcut Start: False
Occlusion: True

A GIT demo (unconstrainted time and space)

Definition: Global instance tracking extends the task of single object tracking by **removing the assumption of continuous motion**, allowing the target to **move freely** between different scenes and camera views. This task aims to **model human dynamic visual capabilities** in more complex and realistic environments.

Global instance tracking is a more human-like task, which allows trackers to align with human visual tracking ability.

S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 45, no. 1, pp. 576–592, 2023.

ACCV 2024

Task Challenge

Discuss the various challenges that arise when performing visual object tracking.

ACCV 2024

Challenges in VOT

- **Challenges are widespread in real world**

➢ **Challenge Overview:** VOT algorithms rely heavily on **appearance and motion information** of the target. When these are disrupted, it leads to errors in predicting the target’s location.

dim light

background clutter

scale / ratio variation

similar object interference

partial occlusion

absent

motion blur

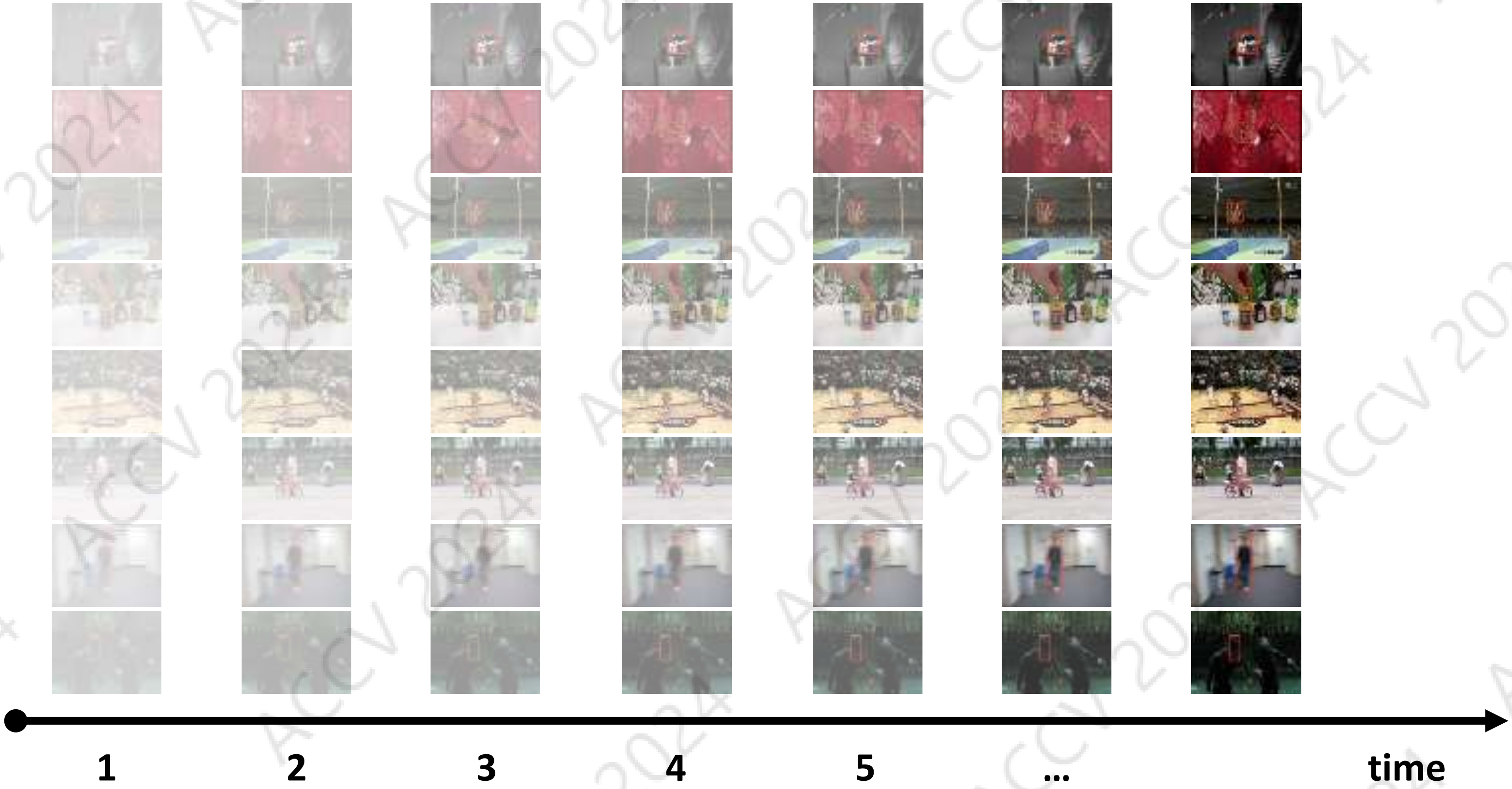
fast motion

ACCV 2024

Challenges in VOT

- **Challenges in real world → Robustness issues**

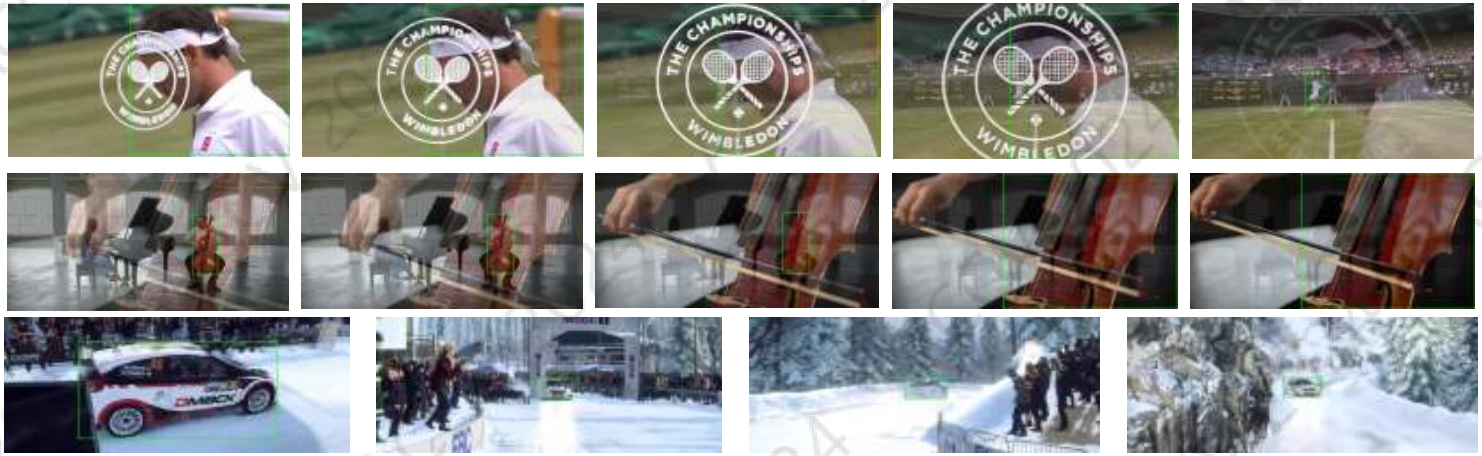
➢ SOT is a sequential decision process. The challenging factors in the environment will **cause errors that continue to accumulate** over time, making it **impossible to achieve robust tracking**.



ACCV 2024

Challenges in VOT

- **Shot-Cut**



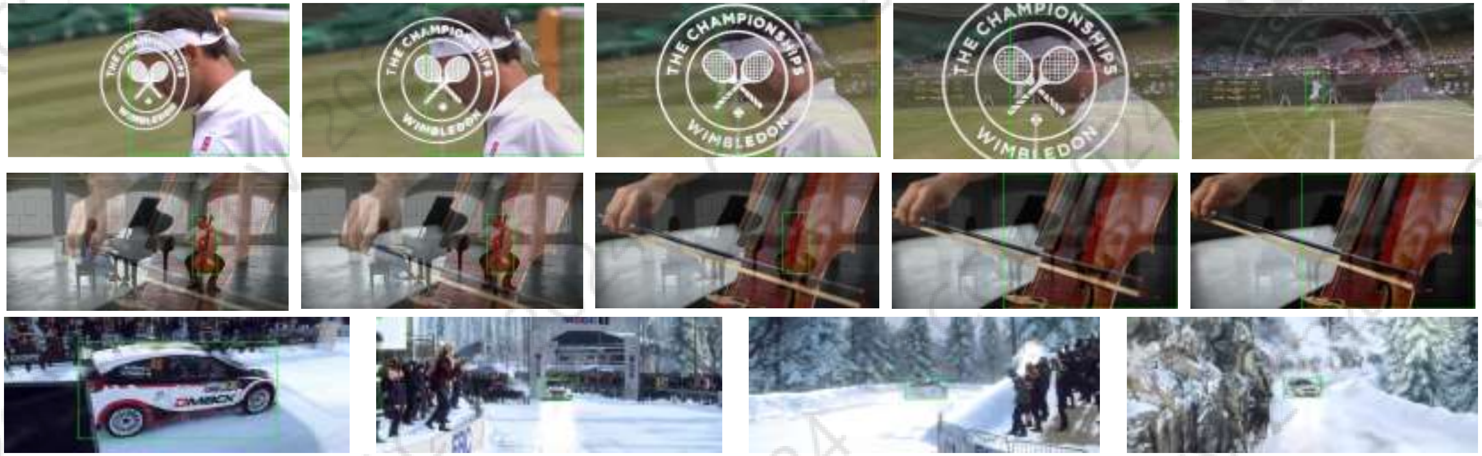
➢ **Appearance Information Disruption:**

- **Scene Transition:**
  - ☐ When there is a shot-cut, **the scene may change entirely**, with the target reappearing in a different context, angle, or lighting.
  - ☐ This sudden transition makes it **difficult for the tracker to maintain the target's visual identity**, as the previously known appearance may no longer be applicable.
- **Change in Target Appearance:** The target may also **look different** after the shot-cut due to changes in camera angle or distance, which **disrupts the consistency of visual features** (like size, texture, or shape).

ACCV 2024

Challenges in VOT

- **Shot-Cut**



➢ **Motion Information Disruption:**

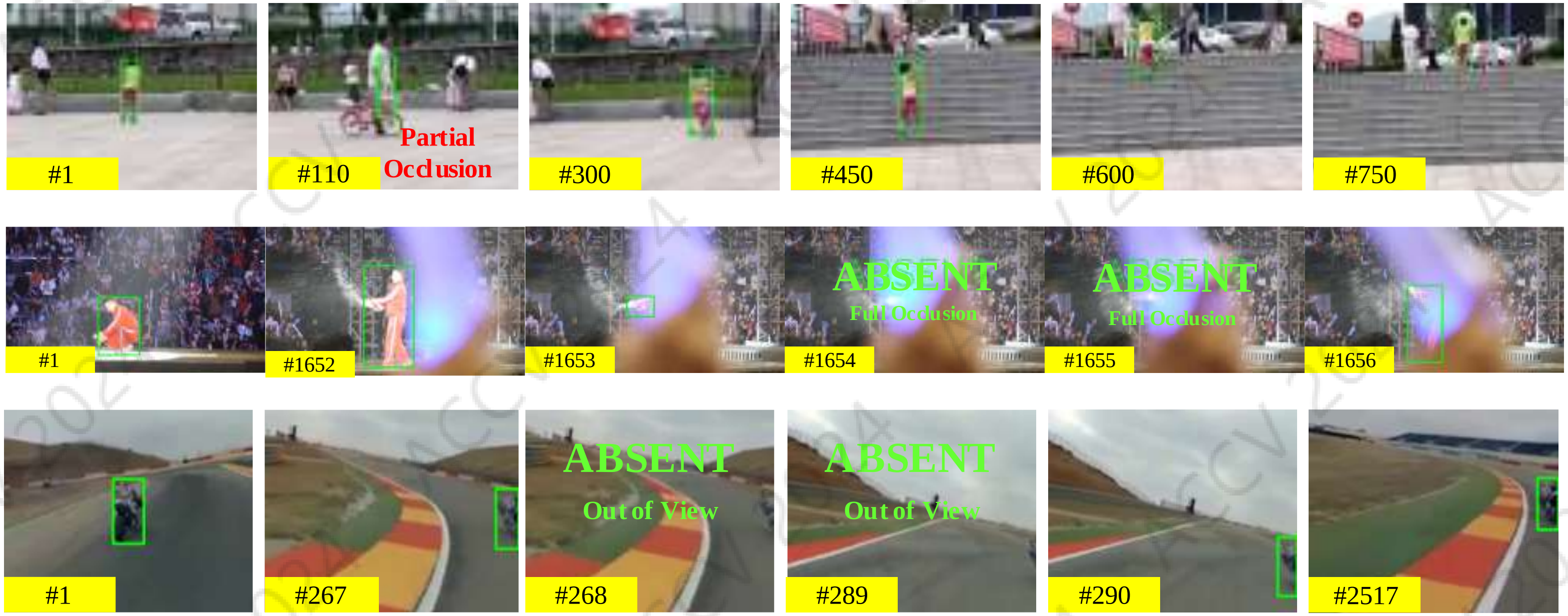
- The natural motion path of the target is interrupted during a shot-cut, leading to **a loss of temporal continuity**.
- The tracker cannot rely on motion data from the previous shot, forcing it to **re-detect the target** in the new frame.

Effect on Tracking:
The tracker must employ **robust re-detection mechanisms** to quickly locate the target in the new shot. Additionally, **context adaptation is necessary** to handle changes in scene and lighting conditions.

ACCV 2024

Challenges in VOT

- **Occlusion and Target Disappearance**


#1
#110
Partial Occlusion
#300
#450
#600
#750
#1
#1652
#1653
#1654
ABSENT
Full Occlusion
#1655
ABSENT
Full Occlusion
#1656
#1
#267
#268
ABSENT
Out of View
#289
ABSENT
Out of View
#290
#2517

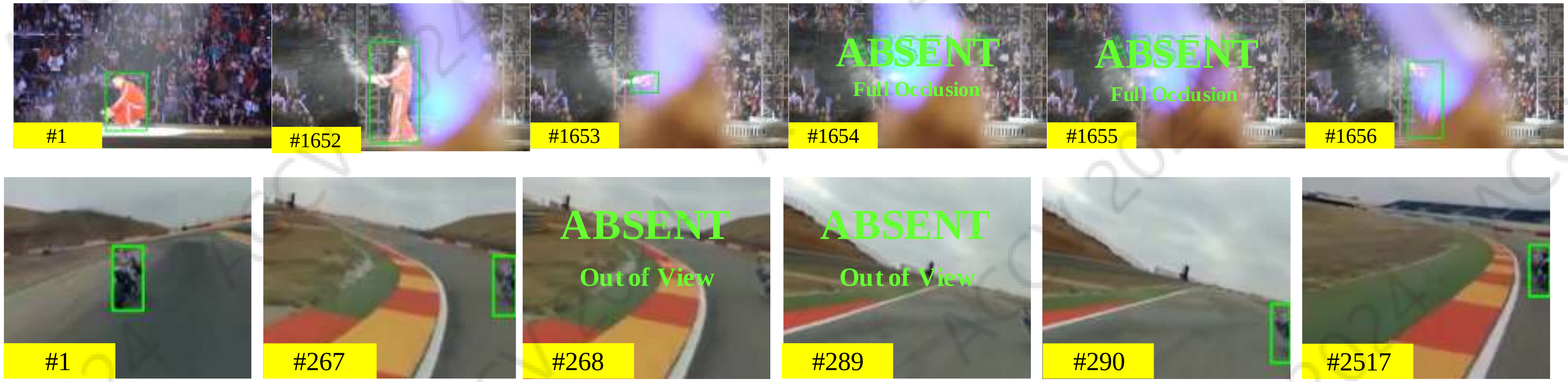
➢ Appearance Information Disruption:

- **Partial Occlusion:** When the target is partially blocked by another object, the algorithm **loses crucial visual information** like texture or color, making it **harder to maintain a precise appearance model**.
- **Full Occlusion:** If the target is completely occluded, the tracker **loses all appearance data**, forcing the algorithm to **rely on motion models or prediction** until the target reappears.

ACCV 2024

Challenges in VOT

- ## Occlusion and Target Disappearance


#1
#1652
#1653
ABSENT
Full Occlusion
#1654
ABSENT
Full Occlusion
#1655
#1656
#1
#267
ABSENT
Out of View
#268
ABSENT
Out of View
#289
#290
#2517

➢ **Motion Information Disruption**:

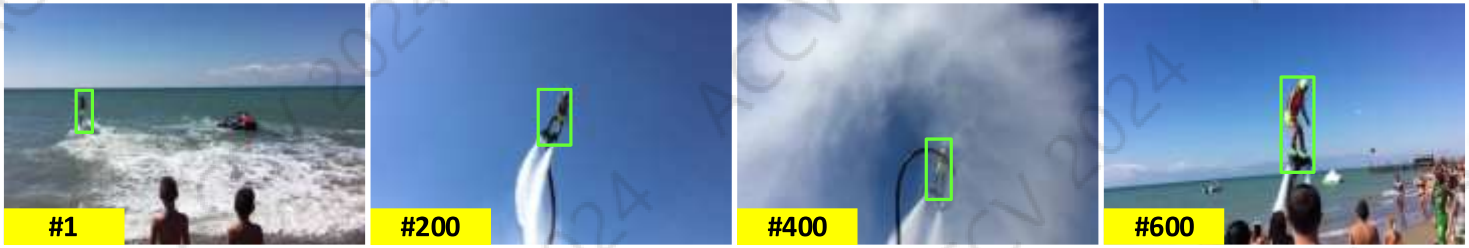
- **Occlusion Causes Loss of Motion Data:** When the target is occluded, the **motion data becomes unavailable or unreliable**, making it **difficult to predict the target's future location.**
- **Target Disappearance:** When the target leaves the field of view or remains fully occluded for an extended period, the algorithm **must handle re-detection**. If it fails, the tracker **may lose the target permanently**.

Effect on Tracking:
The tracker **must adapt to occlusions by predicting the target's likely path using motion models**. Once the target reappears, the tracker should **quickly re-detect it to avoid losing track**.

ACCV 2024

Challenges in VOT

- **Lighting Changes**



➢ **Appearance Information Disruption:**

- Changes in lighting (e.g., moving from a brightly lit area to a shaded region) can **drastically alter the target’s appearance**. These changes may cause the target’s color, texture, or overall brightness to differ significantly from its original appearance.

➢ **Motion Information Disruption:**

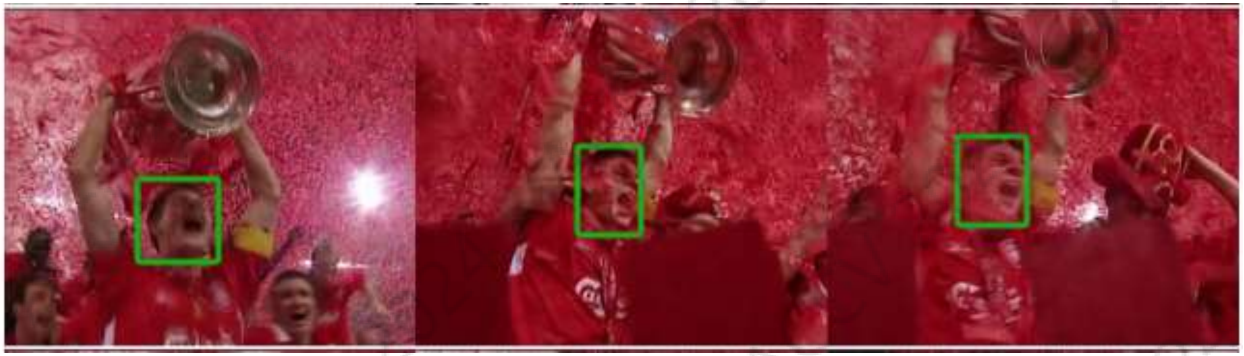
- Poor lighting can **obscure motion information**, making it more challenging for the algorithm to correctly interpret the speed and direction of the target's movement.

Effect on Tracking:
Algorithms must **incorporate adaptive lighting models to handle drastic changes**. Inconsistent lighting can confuse the tracker, leading to incorrect target predictions.

ACCV 2024

Challenges in VOT

- **Background Clutter**



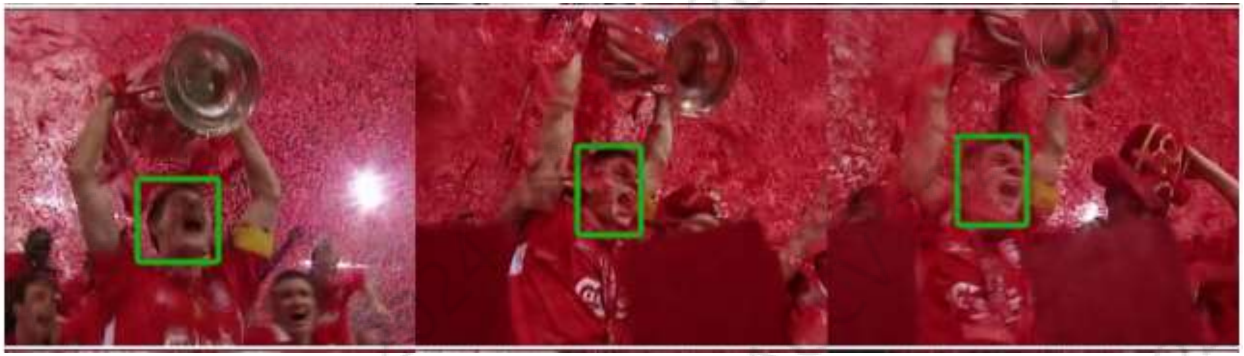
➢ Appearance Information Disruption:

- **Visual Similarity to Background:**
  - ☐ When the target's appearance (e.g., color, texture) is similar to background elements, the tracker may **struggle to differentiate the target from its surroundings**.
  - ☐ This makes it difficult to maintain a clear distinction between the target and background.
- **Distraction by Non-Target Objects:**
  - ☐ In a cluttered background, there **may be many objects that distract the tracker**, especially if these objects have similar visual features.
  - ☐ This confusion can **lead to the tracker locking onto the wrong object or losing the target**.

ACCV 2024

Challenges in VOT

- **Background Clutter**



➢ **Motion Information Disruption**:

- **Interference from Moving Background Elements:**
  - ☐ In dynamic environments (e.g., busy streets or crowded places), **background objects in motion can confuse the tracker**.
  - ☐ The movement of background elements can be misinterpreted as the movement of the target, **leading to errors in motion prediction.**
- **Occlusions by Background Objects:** In some cases, cluttered backgrounds **may temporarily occlude the target**, making it harder for the tracker to estimate the correct motion path.

Effect on Tracking:
Robust appearance models and **motion prediction techniques** are required to distinguish the target from the background in cluttered environments.

ACCV 2024

Challenges in VOT

- **Fast Motion**

Rai Sport + HD
Differita
VOLLEY: CASAL MAGGIORE - SCANDICCI
QUESTA SERA ORE 20:20
demo 1

➢ **Appearance Information Disruption:**

- **Motion Blur:** When the target moves quickly, **motion blur may occur**, causing the visual details (e.g., texture, edges) of the target to become indistinct. This makes it **difficult to maintain a consistent appearance model** of the target.
- **Loss of Visual Features:** In cases of extreme speed, the target may move across frames too quickly, resulting in **a loss of critical visual features** such as shape or color, which the tracker relies on for identification.

ACCV 2024

Challenges in VOT

ACCV

- **Fast Motion**

Rai Sport + HD
Differita
VOLLEY: CASAL MAGGIORE - SCANDICCI
QUESTA SERA ORE 20:20
demo 1

➢ **Motion Information Disruption**:

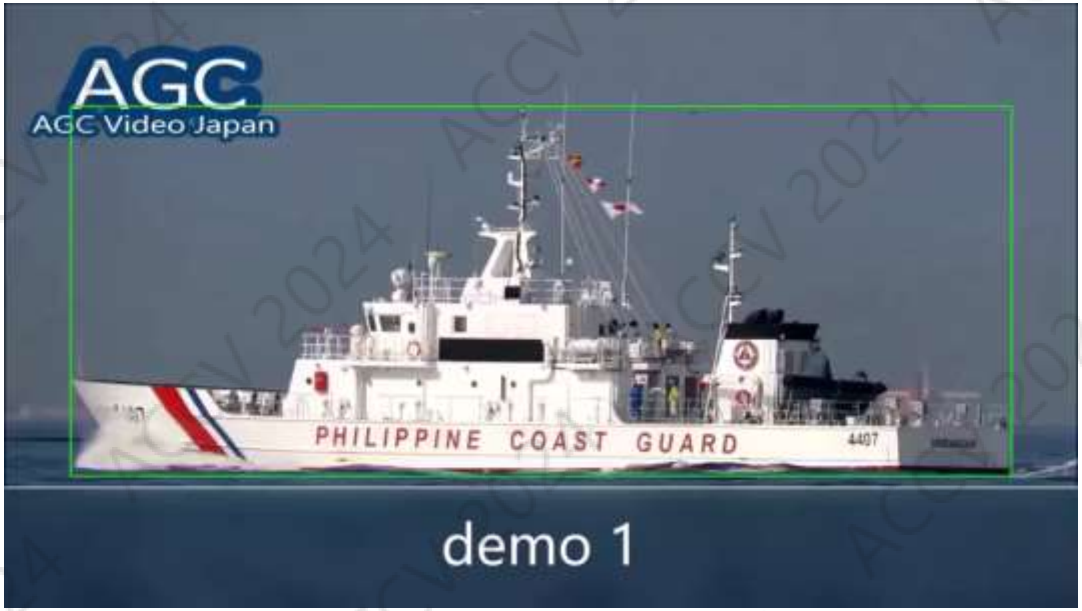
- **High-Speed Movement:** Fast motion makes it **difficult to predict the target's movement accurately**. Traditional motion models may fail to keep up with the rapid changes in the target's position, leading to poor tracking performance.
- **Limited Search Window:** Tracking algorithms often use **a defined search window** around the predicted position. If the target moves too fast, **it may exit the search window**, and the tracker may fail to locate it in the next frame.

Effect on Tracking:
Algorithms **should have more advanced motion models** that can handle sudden and rapid changes in speed and direction.

ACCV 2024

Challenges in VOT

- **Special Scale & Special Ratio**



➢ **Special Scale (Small or Large Targets):**

- **Small Targets:** When tracking small objects, the algorithm may **struggle to capture enough visual detail**, resulting in poor localization accuracy. Small-scale objects have **fewer distinguishable features**, making them harder for the tracker to differentiate from the background.
- **Large Targets:** Conversely, large targets may **exceed the camera's field of view**, resulting in partial occlusion. The tracker must **handle incomplete information**, often leading to inaccuracies in bounding box adjustments.

ACCV 2024

Challenges in VOT

- **Special Scale & Special Ratio**

➢ Special Ratio (Unusual Aspect Ratios):

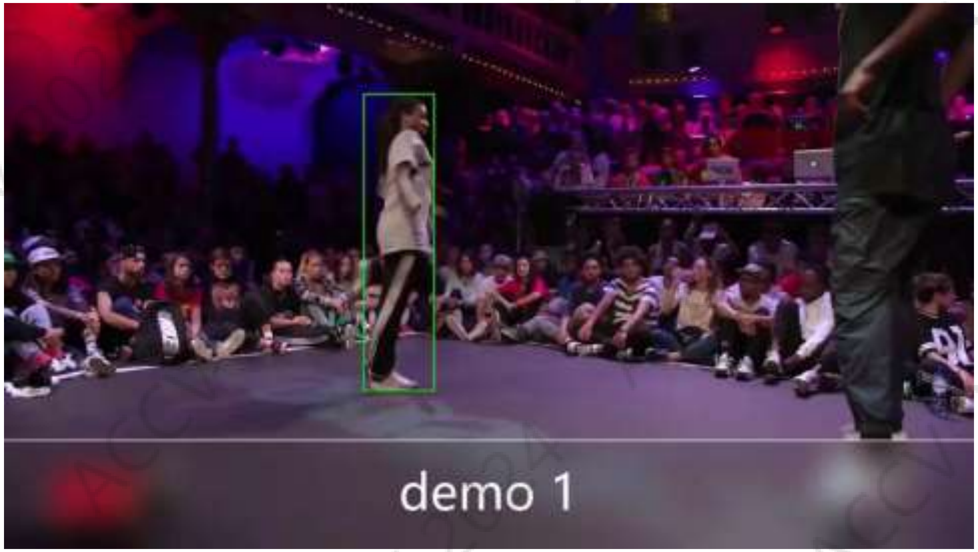
- **Tall or Wide Targets:** Objects with extreme aspect ratios (e.g., very tall or wide) challenge the tracker's ability to accurately fit bounding boxes. Standard tracking models often **struggle with highly elongated objects**, leading to **misalignment of the predicted bounding box with the actual target**.
- **Inconsistent Bounding Box Fitting:** The algorithm's reliance on intersection-over-union (IoU) measures means that special ratio targets often result in poor performance when **the bounding box cannot closely match the target's shape**.

Effect on Tracking:
More flexible and adaptive bounding box models are required to improve tracking accuracy for objects with special scale and ratio characteristics.

ACCV 2024

Challenges in VOT

- **Scale Variation & Ratio Variation**



➢ **Scale Variation:**

- **Dynamic Size Changes**: The target's size in the video frame changes as the relative distance between the target and the camera changes. This **leads to fluctuating scale**, making it difficult for the tracker to maintain accurate predictions.
- **Foreground Feature Alterations:** As the target becomes larger or smaller, key visual features (such as edges, textures) may either become more detailed or be reduced in clarity, **complicating feature extraction for the tracking algorithm.**

Challenges in VOT

- **Scale Variation & Ratio Variation**

➢ **Ratio Variation:**

- **Shape Alterations:** The target's aspect ratio may shift due to rotations or perspective changes, altering the shape of the object in the frame. This **requires the tracker to adjust its bounding box to fit the new shape**, which can be challenging if the ratio changes are extreme.
- **Bounding Box Fitting Issues:** When the aspect ratio changes dramatically, the tracker **may struggle to fit a precise bounding box**, especially in cases where the target becomes highly elongated or compressed.

Effect on Tracking:
More adaptive models capable of **real-time adjustments to both scale and ratio changes** are essential to improve tracking robustness in dynamic scenes.

ACCV 2024

Conclusion

- **Comprehensive Understanding of the Evaluation Task:**
  - This section introduces visual object tracking (VOT) from two key perspectives—**task definitions** and **task challenges.**
  - The goal is to help researchers fully grasp the evaluation task and lay the foundation for intelligent assessment.
- **Constraints in Task Definitions:**
  - The **inherent constraints** in task definitions **reflect the characteristics** of the tracking task.
  - **Changes in these constraints will shift the focus of the evaluation**. For example, a change in the duration of tracking (e.g., short-term vs. long-term) could change the way algorithms are evaluated.

ACCV 2024

Conclusion

- **Task Challenges Represent Difficulties:**
  - The challenging factors in SOT represent the **primary difficulties**.
  - A deep understanding of these factors is crucial for researchers to **accurately identify the performance bottlenecks** of tracking algorithms.
- **Importance of Designing Evaluation Environments:**
  - Only by understanding the specific task challenges, such as occlusion, fast motion, and scale variation, can researchers **design the right evaluation environments** that test an algorithm's real capabilities.
  - Proper evaluation ensures that the algorithm's weaknesses are uncovered, **enabling improvement and refinement**.

ACCV 2024

CONTENTS

- Introduction
- Part 1. Visual Object Tracking Task
- **Part 2. Experimental Environments**
  - General Datasets
  - Specialized Datasets
  - Competition Datasets
- Part 3. Algorithms and Traditional Machine-Machine Comparisons
- Part 4. Human Visual Abilities and Visual Turing Test
- Trends and Future Directions

Scan to download this tutorial PPT

ACCV 2024

General Datasets

General datasets are designed to test the performance of algorithms under a variety of conditions.

ACCV 2024

General Datasets: Small-scale

- **OTB50 (2013) & OTB100 (2015)**

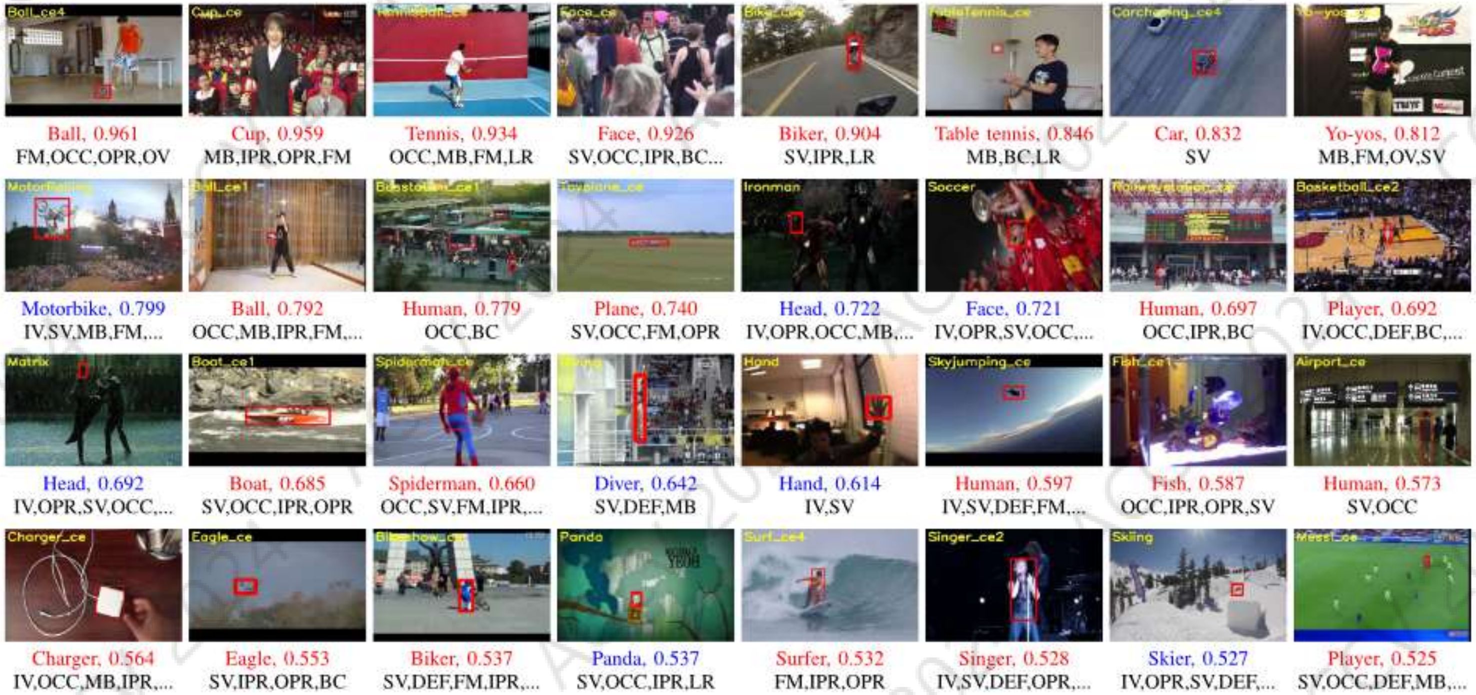
➢ OTB50 is **one of the earliest benchmarks** designed specifically for evaluating SOT algorithms. OTB100 extended OTB50 by including more tracking sequences and covering a broader set of tracking scenarios. OTB provides high-precision annotations using horizontal bounding boxes and includes a variety of tracking challenges such as occlusion, fast motion, and scale variation.

- **Standardization:** OTB helped **standardize SOT evaluations** by offering a set of predefined benchmarks, enabling researchers to compare their algorithms on a unified platform.
- **Challenge Annotations:** OTB is annotated with **multiple challenge factors**, making it a comprehensive evaluation platform for early tracking algorithms.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

General Datasets: Small-scale

- **TColor-128 (2015)**


Ball, 0.961
FM,OCC,OPR,OV
Cup, 0.959
MB,IPR,OPR,FM
Tennis, 0.934
OCC,MB,FM,LR
Face, 0.926
SV,OCC,IPR,BC...
Biker, 0.904
SV,IPR,LR
Table tennis, 0.846
MB,BC,LR
Car, 0.832
SV
Yo-yos, 0.812
MB,FM,OV,SV
Motorbike, 0.799
IV,SV,MB,FM,...
Ball, 0.792
OCC,MB,IPR,FM,...
Human, 0.779
OCC,BC
Plane, 0.740
SV,OCC,FM,OPR
Head, 0.722
IV,OPR,OCC,MB,...
Face, 0.721
IV,OPR,SV,OCC,...
Human, 0.697
OCC,IPR,BC
Player, 0.692
IV,OCC,DEF,BC,...
Head, 0.692
IV,OPR,SV,OCC,...
Boat, 0.685
SV,OCC,IPR,OPR
Spiderman, 0.660
OCC,SV,FM,IPR,...
Diver, 0.642
SV,DEF,MB
Hand, 0.614
IV,SV
Human, 0.597
IV,SV,DEF,FM,...
Fish, 0.587
OCC,IPR,OPR,SV
Human, 0.573
SV,OCC
Charger, 0.564
IV,OCC,MB,IPR,...
Eagle, 0.553
SV,IPR,OPR,BC
Biker, 0.537
SV,DEF,FM,IPR,...
Panda, 0.537
SV,OCC,IPR,LR
Surfer, 0.532
FM,IPR,OPR
Singer, 0.528
IV,SV,DEF,OPR,...
Skier, 0.527
IV,OPR,SV,DEF,...
Player, 0.525
SV,OCC,DEF,MB,...

- ➢ TColor-128 aims to **evaluate the role of color features** in SOT, particularly in distinguishing targets from complex backgrounds.
  - **Focus on Color Features:** Unlike datasets that include grayscale sequences, TColor-128 exclusively **includes color video sequences**, enabling the evaluation of algorithms that rely on color information to differentiate targets.

Liang P, Blasch E, Ling H. Encoding color information for visual tracking: Algorithms and benchmark[J]. IEEE transactions on image processing, 2015, 24(12): 5630-5644.

ACCV 2024

General Datasets: Small-scale

- **Limitations**

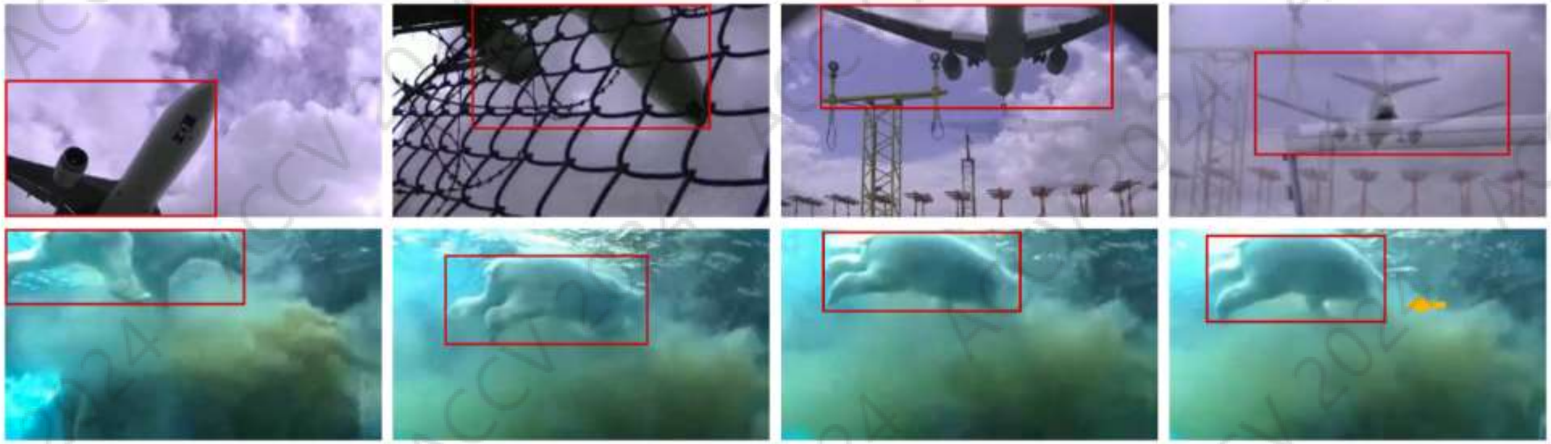
➢ While small-scale datasets, like the ones we've discussed, provide valuable benchmarks, they also come with certain limitations that affect the performance of deep learning models.

- **Data Volume Constraints**: Deep learning models require a **large amount of labeled data** to achieve optimal performance. Many traditional SOT datasets are limited in size, which **constrains the ability of deep learning models to generalize** across diverse scenarios and environments.
- **Poor Generalization**: Models trained on smaller datasets may suffer from **poor generalization** when tested in more complex real-world environments, especially when faced with **unseen target types** or **challenging conditions** like fast motion or occlusion.
- **Lack of Diversity**: Small-scale datasets often **lack diversity in both target types and tracking conditions**, which limits the **robustness** of tracking algorithms. This makes it harder for these models to handle new or unexpected scenarios.

More large-scale tracking datasets with dynamic objects and varied tracking challenges are necessary to enhance the performance of deep learning-based tracking models.

General Datasets: Large-scale

- **ImageNet-VID (2015) & YouTube-BB (2017)**



- Two datasets from video object detection task:
  - ImageNet-VID contains 5,400 video sequences with annotations for one or more moving objects.
  - YouTube-BB includes 380,000 YouTube video sequences with 5.4 million frames annotated at a rate of 1 Hz.
- Despite its large size, these datasets focus on **a small number of object categories** and **includes static objects**, which limits their utility for training dynamic object tracking models.
- To overcome the limitations of small datasets and the static nature of larger datasets like ImageNet-VID and YouTube-BB, there is a need for **more diverse and dynamic large-scale datasets** designed for SOT task.

Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115: 211-252.

Real E, Shlens J, Mazzocchi S, et al. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5296-5305.

ACCV 2024

General Datasets: Large-scale STT

- **TrackingNet (2018)**

➢ TrackingNet is **one of the largest datasets for short-term tracking**, designed to support deep learning-based tracking models.

  - **Filtered for Quality:** The dataset **filters out static objects and noisy segments** from YouTube-BB, focusing on moving objects that are relevant for tracking tasks.
  - **Combines Manual and Automated Annotations:** Uses discriminative correlation filter (DCF) to **automate annotation**, **combined with manual annotations** for greater precision.

Muller M, Bibi A, Giancola S, et al. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 300-317.

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ **SOT Definition:** Provides only the initial position of a moving object, and continuously locates it in a video sequence.

➢ **SOT Characteristics:**

- **Sequential decision:** locating the target with the help of previous frames
- **Category agnostic:** without any assumption about the target category (open-set setting)
- **Instance-level prediction:** need to distinguish the target from others (including objects in the same category)

There are a large number of unknown target categories in the real environment. → **Generalization Challenges**

L. Huang, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.*

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ **Motivation:** Limitations of the existing experimental environment in terms of generalization:

- Weak diversity, non-universal scenarios, **poor generalization ability** (training and test categories completely overlap and have the same distribution)

LaSOT (CVPR'19)

- 70 object categories
- **Training and testing categories completely overlap and have consistent distribution**

TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild

TrackingNet (ECCV'19)

- 22 object categories
- **Training and testing categories completely overlap and have consistent distribution**

ACCV 2024

L. Huang*, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ A large-scale benchmark that covers a wide range of **natural and artificial object categories** and **motion forms** and follows the **open set evaluation protocol.**

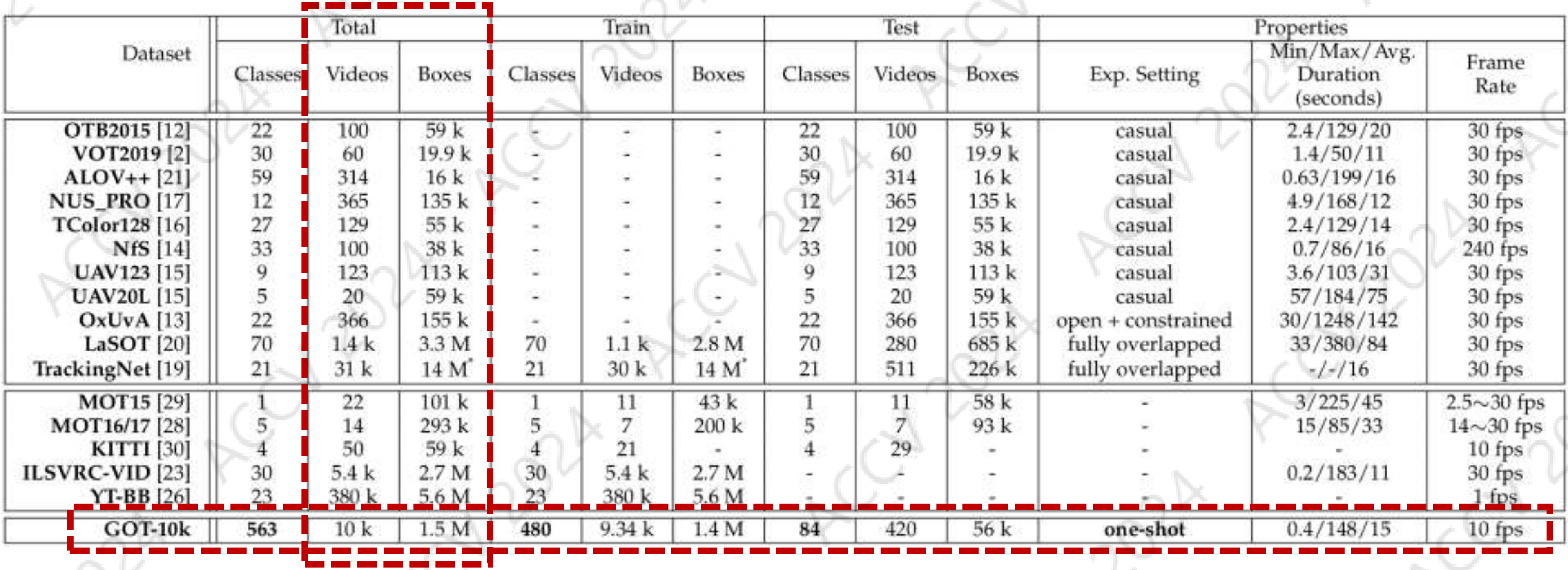
L. Huang*, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.

ACCV 2024

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ Large-scale, **unified** training, validation, and test sets.



Dataset	Total			Train			Test			Properties		
	Classes	Videos	Boxes	Classes	Videos	Boxes	Classes	Videos	Boxes	Exp. Setting	Min/Max/Avg. Duration (seconds)	Frame Rate
OTB2015 [12]	22	100	59 k	-	-	-	22	100	59 k	casual	2.4/129/20	30 fps
VOT2019 [2]	30	60	19.9 k	-	-	-	30	60	19.9 k	casual	1.4/50/11	30 fps
ALOV++ [21]	59	314	16 k	-	-	-	59	314	16 k	casual	0.63/199/16	30 fps
NUS_PRO [17]	12	365	135 k	-	-	-	12	365	135 k	casual	4.9/168/12	30 fps
TColor128 [16]	27	129	55 k	-	-	-	27	129	55 k	casual	2.4/129/14	30 fps
NfS [14]	33	100	38 k	-	-	-	33	100	38 k	casual	0.7/86/16	240 fps
UAV123 [15]	9	123	113 k	-	-	-	9	123	113 k	casual	3.6/103/31	30 fps
UAV20L [15]	5	20	59 k	-	-	-	5	20	59 k	casual	57/184/75	30 fps
OxUvA [13]	22	366	155 k	-	-	-	22	366	155 k	open + constrained	30/1248/142	30 fps
LaSOT [20]	70	1.4 k	3.3 M	70	1.1 k	2.8 M	70	280	685 k	fully overlapped	33/380/84	30 fps
TrackingNet [19]	21	31 k	14 M*	21	30 k	14 M*	21	511	226 k	fully overlapped	-/-/16	30 fps
MOT15 [29]	1	22	101 k	1	11	43 k	1	11	58 k	-	3/225/45	2.5~30 fps
MOT16/17 [28]	5	14	293 k	5	7	200 k	5	7	93 k	-	15/85/33	14~30 fps
KITTI [30]	4	50	59 k	4	21	-	4	29	-	-	-	10 fps
ILSVRC-VID [23]	30	5.4 k	2.7 M	30	5.4 k	2.7 M	-	-	-	-	0.2/183/11	30 fps
YT-BB [26]	23	380 k	5.6 M	23	380 k	5.6 M	-	-	-	-	-	1 fps
GOT-10k	**563**	10 k	1.5 M	**480**	9.34 k	1.4 M	**84**	420	56 k	**one-shot**	0.4/148/15	10 fps

10k videos， 1.5M manual annotations

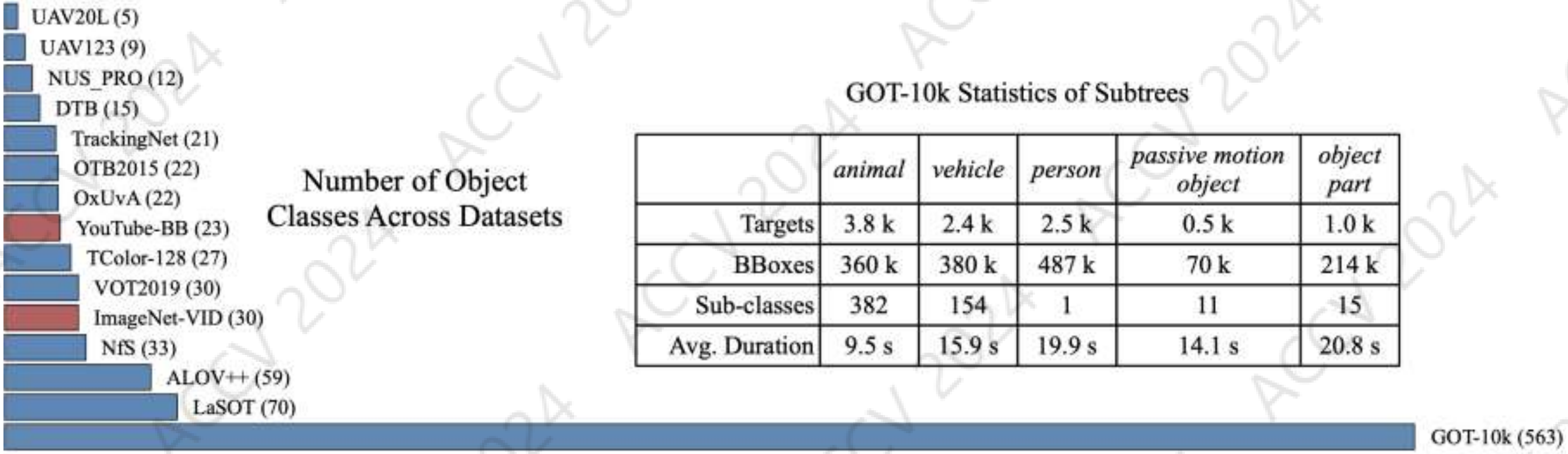
L. Huang, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.*

ACCV 2024

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ 563 types of objects, fully covering common natural and man-made moving objects in **WordNet.**



GOT-10k Statistics of Subtrees

	animal	*vehicle*	*person*	*passive motion object*	*object part*
Targets	3.8 k	2.4 k	2.5 k	0.5 k	1.0 k
BBoxes	360 k	380 k	487 k	70 k	214 k
Sub-classes	382	154	1	11	15
Avg. Duration	9.5 s	15.9 s	19.9 s	14.1 s	20.8 s

The number of object categories (563) is nearly 10 times that of other tracking datasets (in 2019)

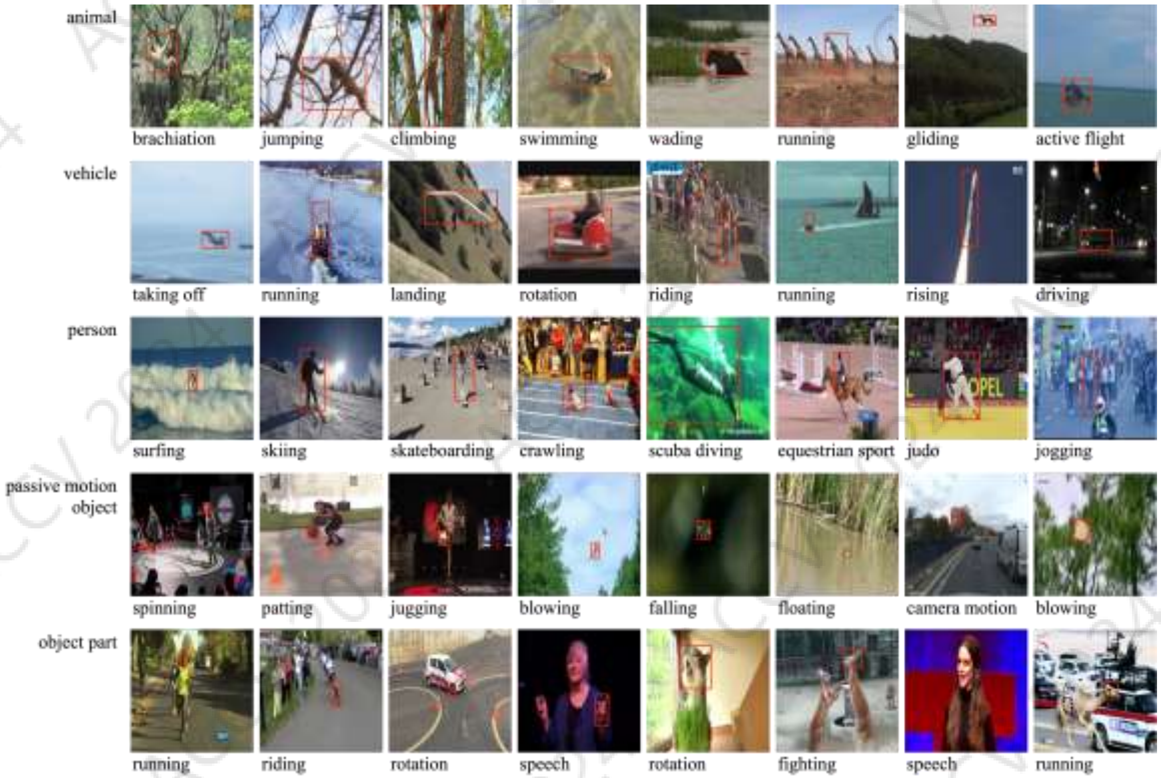
L. Huang, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.*

ACCV 2024

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ 87 types of motion modes, covering a **wide range** of different forms of sports trajectories.


animal
brachiation
jumping
climbing
swimming
wading
running
gliding
active flight
vehicle
taking off
running
landing
rotation
riding
running
rising
driving
person
surfing
skiing
skateboarding
crawling
scuba diving
equestrian sport
judo
jogging
passive motion object
spinning
patting
jugging
blowing
falling
floating
camera motion
blowing
object part
running
riding
rotation
speech
rotation
fighting
speech
running

Object category labels

Motion category labels

Each video sequence includes two labels: object category and motion category.

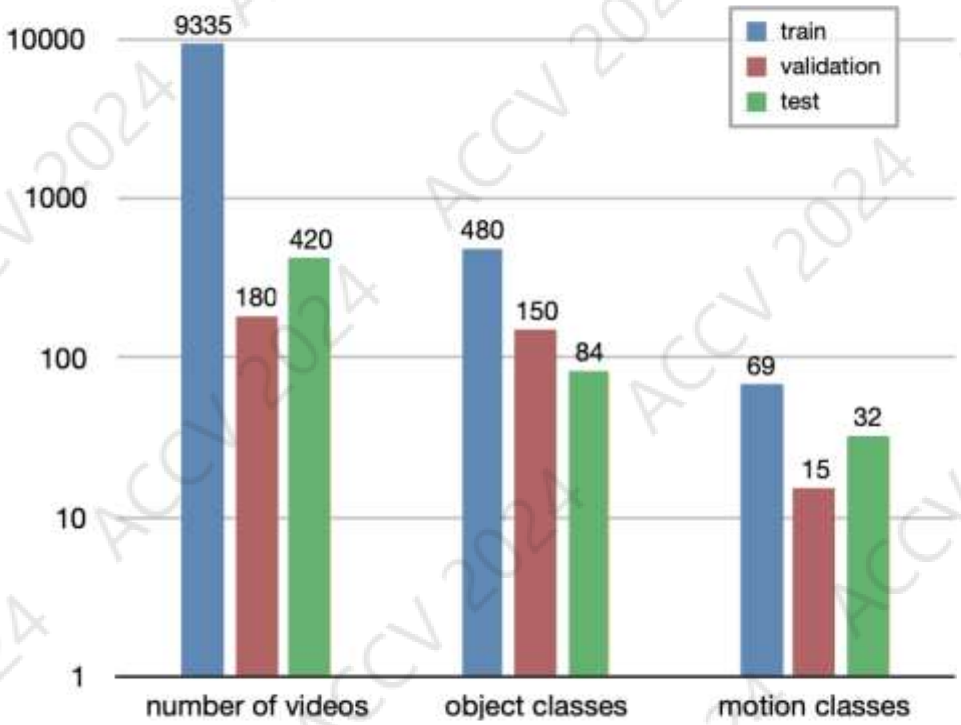
L. Huang, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.*

ACCV 2024

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ **Open set evaluation specification** (training and test categories do not overlap at all) for **generalization ability evaluation**



There is no overlap between training and testing categories, and the algorithm is required to accurately track moving objects of unknown categories.

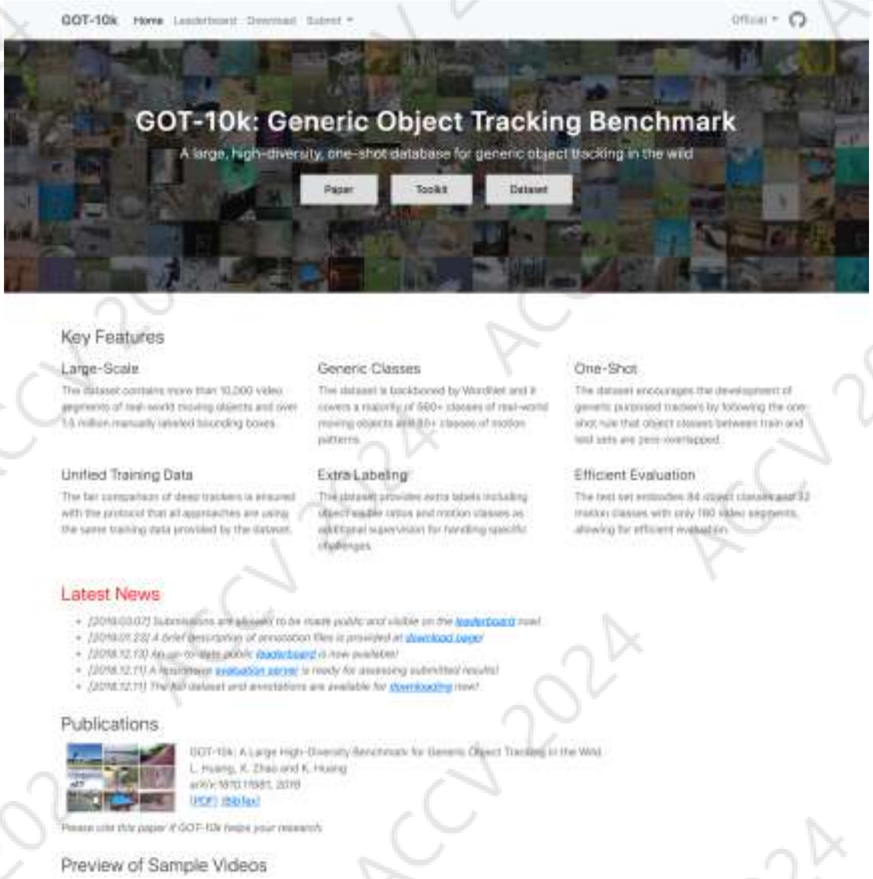
L. Huang*, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.

ACCV 2024

General Datasets: Large-scale STT

- **GOT-10k (2019)**

➢ Complete evaluation platform, **real-time rankings**, open-source toolkits


GOT-10k
Home
Leaderboard
Download
Submit
Official
Paper
Toolkit
Dataset

GOT-10k: Generic Object Tracking Benchmark

A large, high-diversity, one-shot database for generic object tracking in the wild

Key Features

Large-Scale

Generic Classes

One-Shot

Unified Training Data

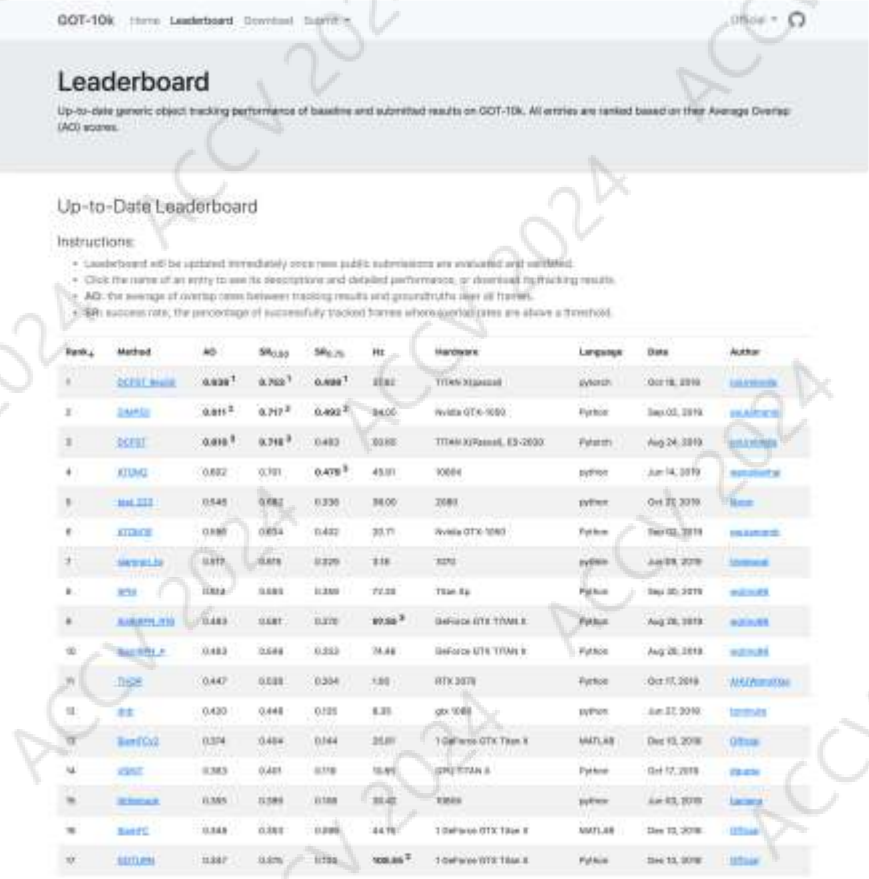
Extra Labeling

Efficient Evaluation

Latest News

Publications

Preview of Sample Videos


GOT-10k
Home
Leaderboard
Download
Submit
Official

Leaderboard

Up-to-date generic object tracking performance of baseline and submitted results on GOT-10k. All entries are ranked based on their Average Overlap (AO) scores.

Up-to-Date Leaderboard

Instructions:

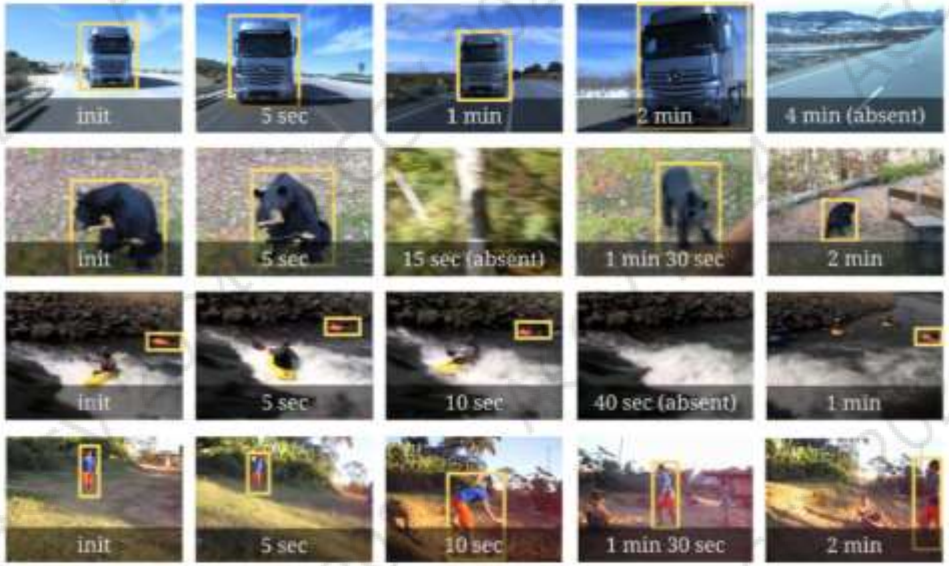
Rank	Method	AO	$SR_{0.50}$	$SR_{0.75}$	Hz	Hardware	Language	Date	Author

http://got-10k.aitestunion.com/

L. Huang*, X. Zhao*, K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1562-1577.

General Datasets: Large-scale LTT

- **OxUvA (2018)**


init
5 sec
1 min
2 min
4 min (absent)
init
5 sec
15 sec (absent)
1 min 30 sec
2 min
init
5 sec
10 sec
40 sec (absent)
1 min
init
5 sec
10 sec
1 min 30 sec
2 min

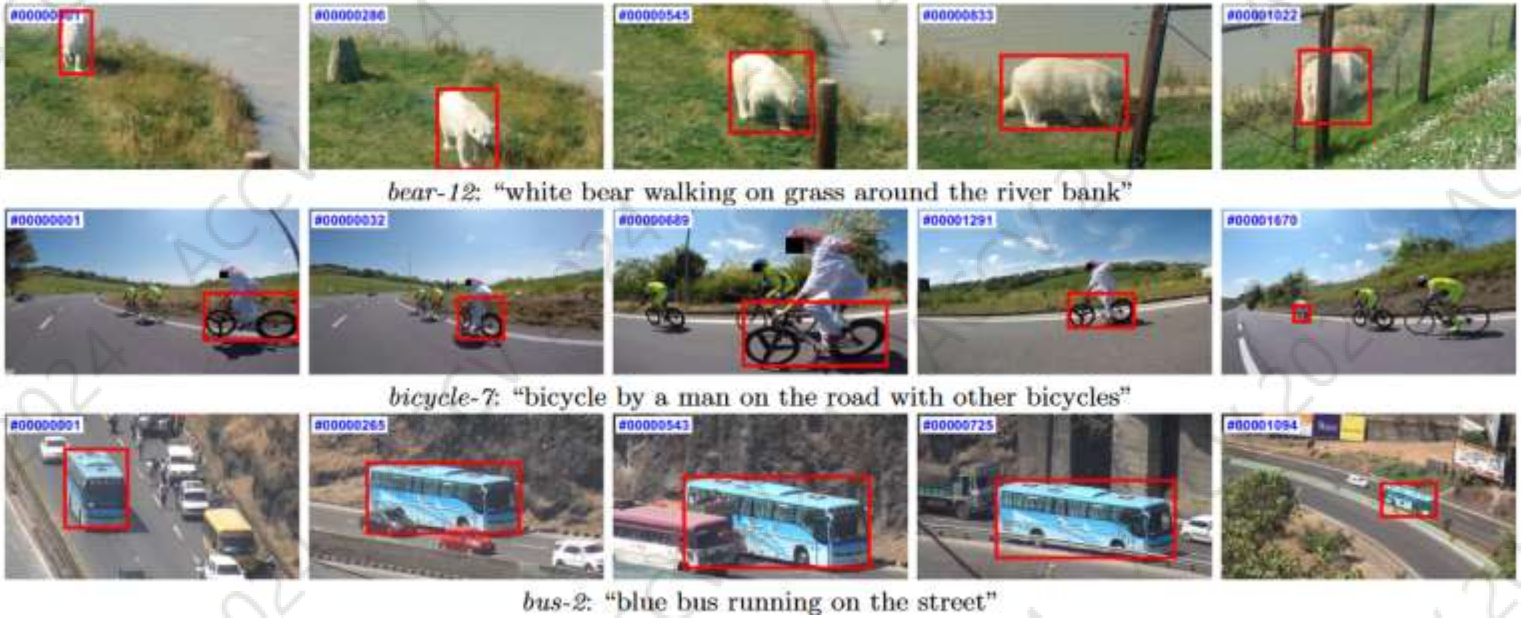
➢ OxUvA is designed specifically for **long-term object tracking**, testing the ability of algorithms to handle **target disappearance and reappearance** across extended video sequences.

  - **Challenging Long-term Evaluations:** This dataset shifts the focus from tracking in consistent, short-term scenarios to more **dynamic and unpredictable** long-term tracking conditions.
  - **Benchmark for Robustness:** OxUvA tests the robustness of tracking algorithms by introducing the challenge of **target disappearance**, making it a critical dataset for evaluating next-generation tracking models.

Valmadre J, Bertinetto L, Henriques J F, et al. Long-term tracking in the wild: A benchmark[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 670-685.

General Datasets: Large-scale LTT

• LaSOT (2019) & LaSOT-ext (2021)



bear-12: "white bear walking on grass around the river bank"

bicycle-7: "bicycle by a man on the road with other bicycles"

bus-2: "blue bus running on the street"

➢ LaSOT is designed to evaluate long-term single object tracking algorithms, with extended sequences to test tracking robustness over long durations.

- **Standard for Long-term Tracking:** LaSOT is one of the largest and most comprehensive datasets for long-term tracking, offering **diverse sequences and challenges** that test algorithms beyond short-term scenarios.
- **Semantic Annotations:** The dataset also provides **semantic annotations** for each sequence, making it useful for **multimodal research** and advanced tracking techniques that require semantic understanding.

Fan H, Lin L, Yang F, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5374-5383.

ACCV 2024

General Datasets: Large-scale GIT

- **VideoCube (2023)**

➢ **Motivation:** Limitations of the existing experimental environment in terms of robustness:

  - The continuous motion assumption limits the experimental environment to **simple scenes with slow motion and a single shot.**

LaSOT: Continuous motion assumption, no shot cuts and scene transitions → Simple environment

*S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

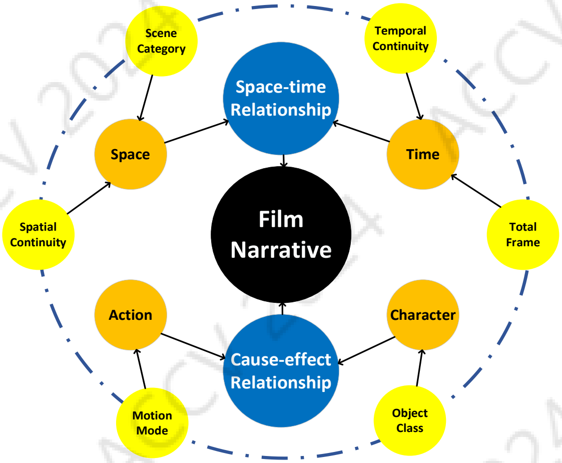
ACCV 2024

General Datasets: Large-scale GIT

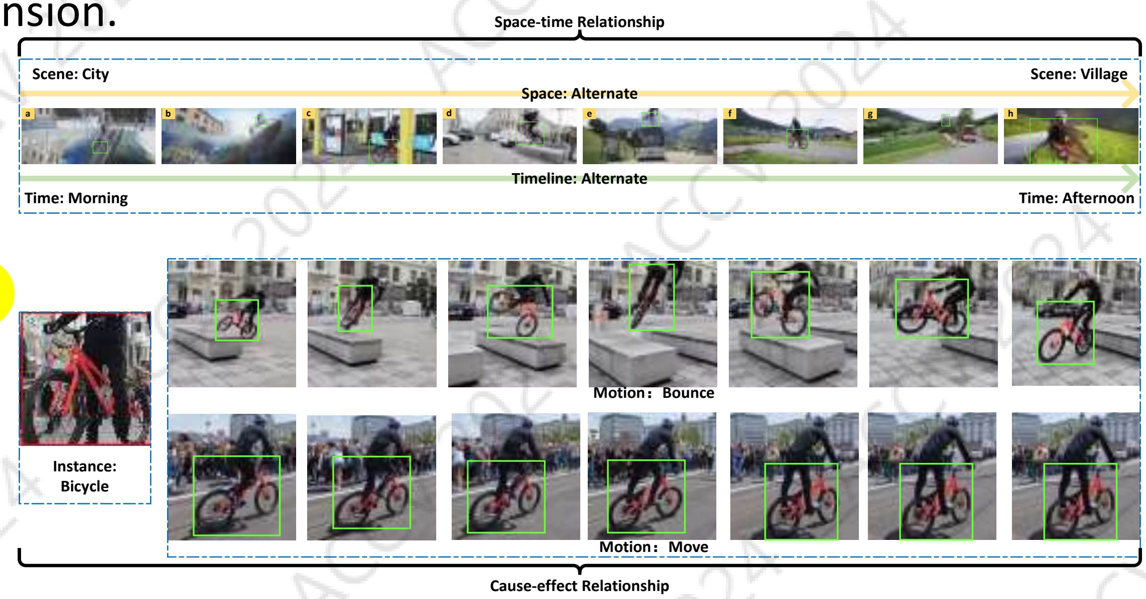
- **VideoCube (2023)**

➢ **Scientific collection principles:**

  - Based on the **narrative theory of film**, the 6D principle is proposed to simulate real scenes.
  - For the first time, **scene categories and spatio-temporal factors** are included in the collection dimension.


Scene Category
Temporal Continuity
Space-time Relationship
Space
Time
Spatial Continuity
Film Narrative
Total Frame
Action
Character
Cause-effect Relationship
Motion Mode
Object Class

6D Principle


Space-time Relationship
Scene: City
Scene: Village
Space: Alternate
a
b
c
d
e
f
g
h
Timeline: Alternate
Time: Morning
Time: Afternoon
Instance: Bicycle
Motion：Bounce
Motion：Move
Cause-effect Relationship

Collection Dimension

Film narrative is a chain of events in **cause-effect relationship** occurring in **space and time**.
(Bordwell & Thompson 2004, Film art : an introduction)

*S. Hu, X. Zhao#, L. Huang, et al., “Global instance tracking: Locating target more like humans,” IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

ACCV 2024

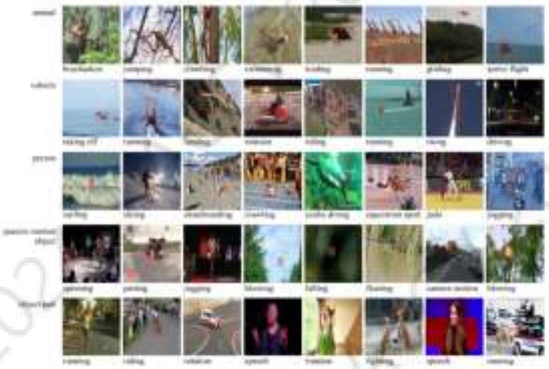
General Datasets: Large-scale GIT

- **VideoCube (2023)**

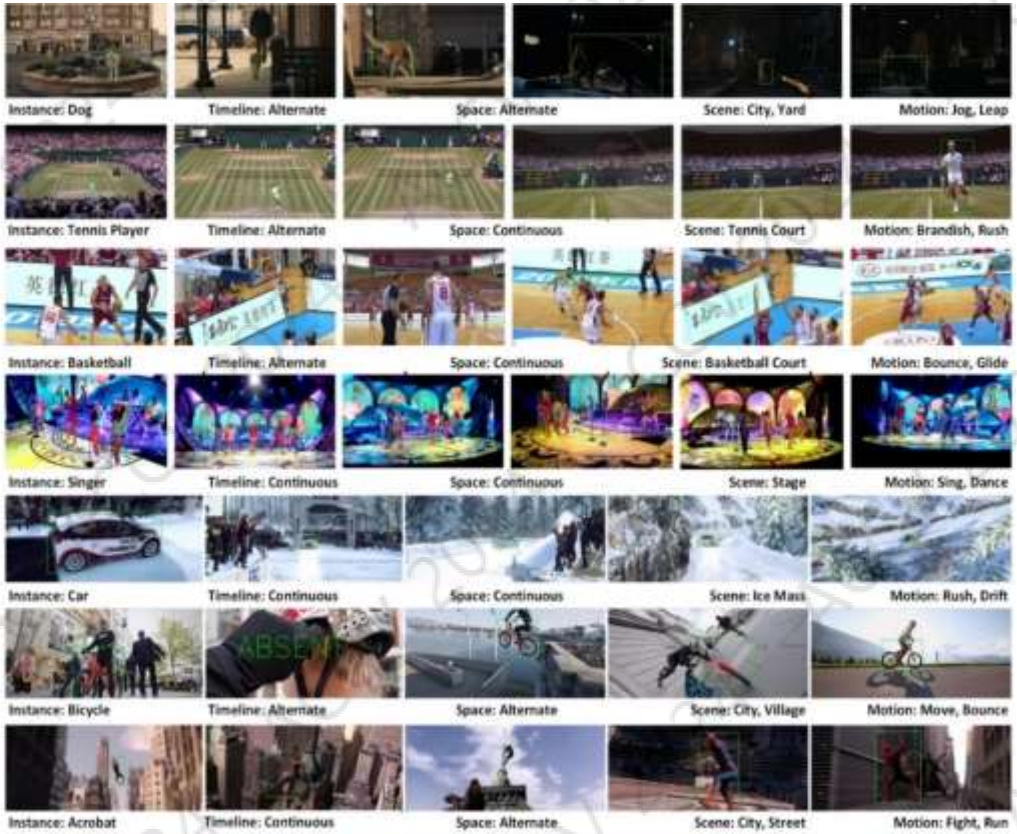
➢ **Scientific collection principles:**

  - Compared with existing datasets, VideoCube has **richer content.**

LaSOT: Object Classes



GOT-10k: Object categories + motion modes


Instance: Dog
Timeline: Alternate
Space: Alternate
Scene: City, Yard
Motion: Jog, Leap
Instance: Tennis Player
Timeline: Alternate
Space: Continuous
Scene: Tennis Court
Motion: Brandish, Rush
Instance: Basketball
Timeline: Alternate
Space: Continuous
Scene: Basketball Court
Motion: Bounce, Glide
Instance: Singer
Timeline: Continuous
Space: Continuous
Scene: Stage
Motion: Sing, Dance
Instance: Car
Timeline: Continuous
Space: Continuous
Scene: Ice Mass
Motion: Rush, Drift
Instance: Bicycle
Timeline: Alternate
Space: Alternate
Scene: City, Village
Motion: Move, Bounce
Instance: Acrobat
Timeline: Continuous
Space: Alternate
Scene: City, Street
Motion: Fight, Run

VideoCube: Target category + motion mode + scene category + spatiotemporal continuity

*S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

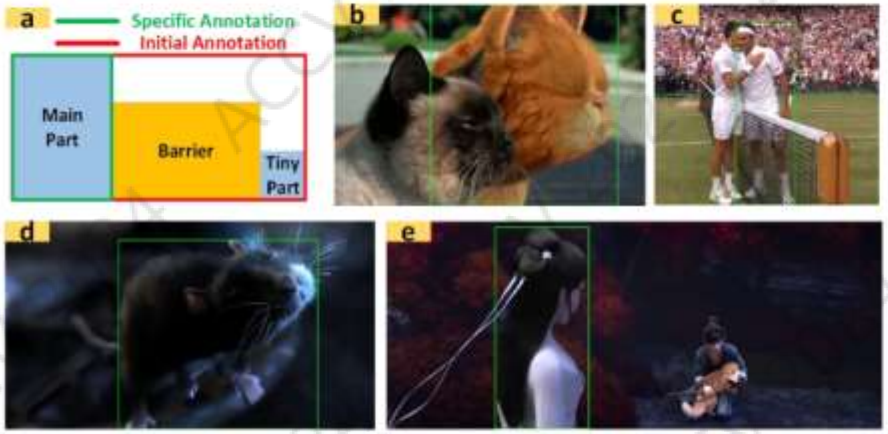
ACCV 2024

General Datasets: Large-scale GIT

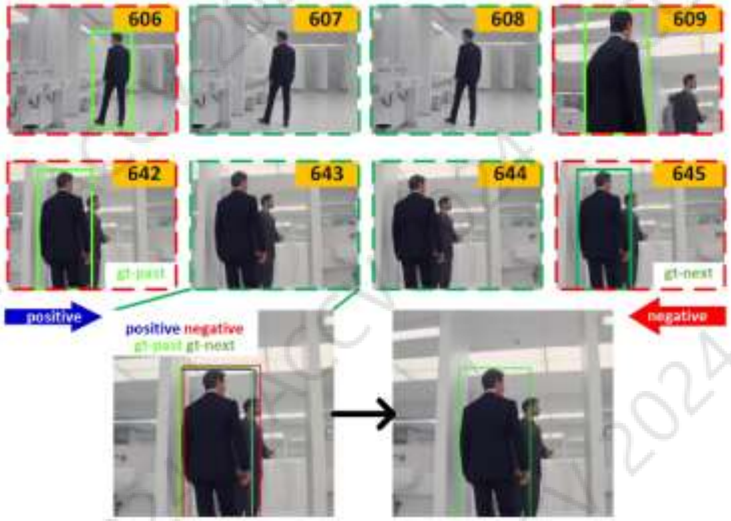
- **VideoCube (2023)**

➢ **High-precision labeling:**

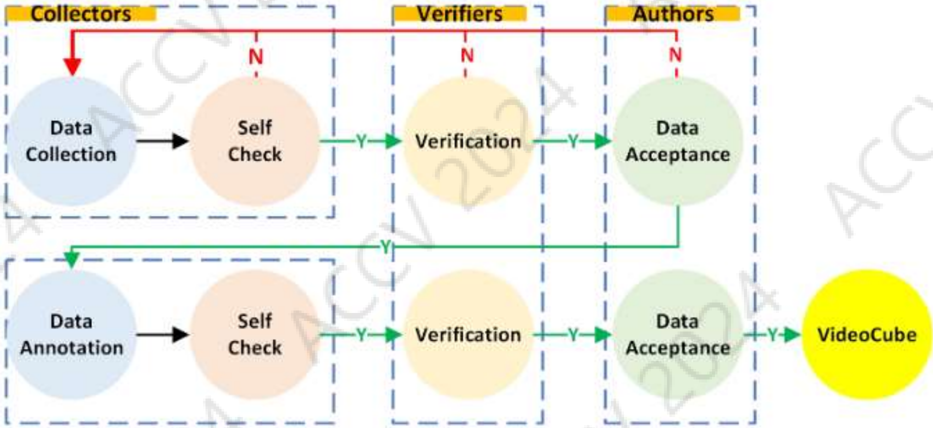
- Standardized labeling criteria + strict review process → **Improve data quality**


a
Specific Annotation
Initial Annotation
Main Part
Barrier
Tiny Part
b
c
d
e

Specific rules in annotations


606
607
608
609
642
643
644
645
gt-past
gt-next
positive
negative
positive negative
gt-past gt-next

Manual & automatic annotations


Collectors
Verifiers
Authors
N
N
N
Data Collection
Self Check
Y
Verification
Y
Data Acceptance
Y
Data Annotation
Self Check
Y
Verification
Y
Data Acceptance
Y
VideoCube

Data checkout process

S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 45, no. 1, pp. 576–592, 2023.

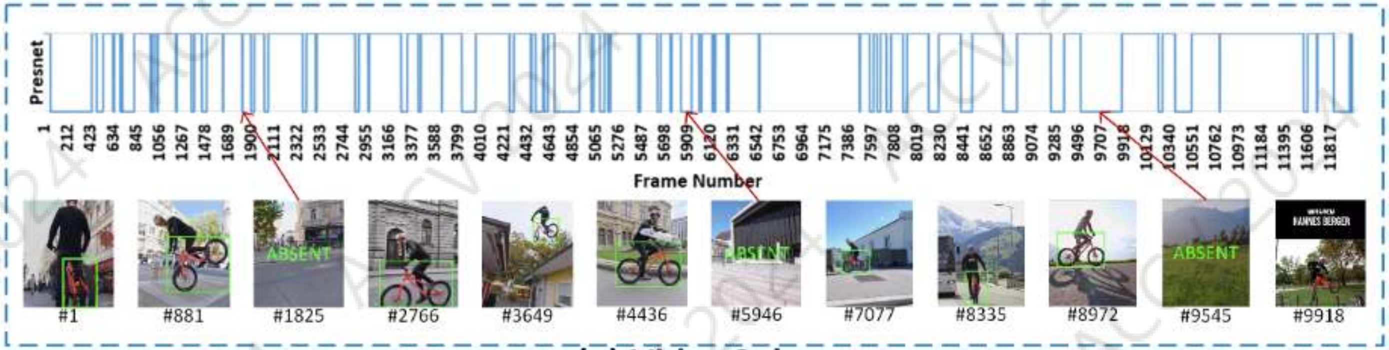
ACCV 2024

General Datasets: Large-scale GIT

- **VideoCube (2023)**

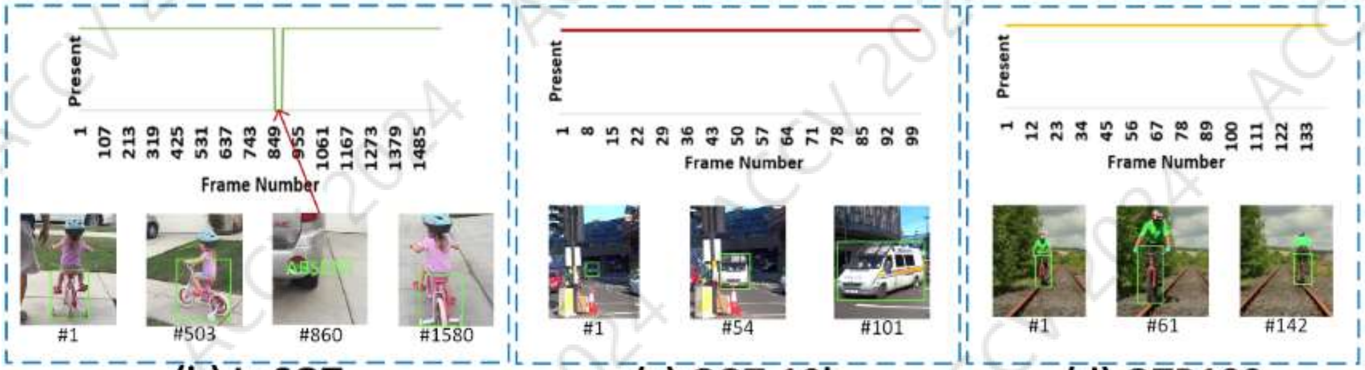
➢ **Large-scale dataset:**

- The duration of a single video segment is much longer than existing tracking datasets and **includes camera switching and scene transitions**.



(a) VideoCube

Long video sequences + rich shot switching and scene transitions



(b) LaSOT **(c) GOT-10k** **(d) OTB100**

Short video sequence + single shot + fixed scene

S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 45, no. 1, pp. 576–592, 2023.

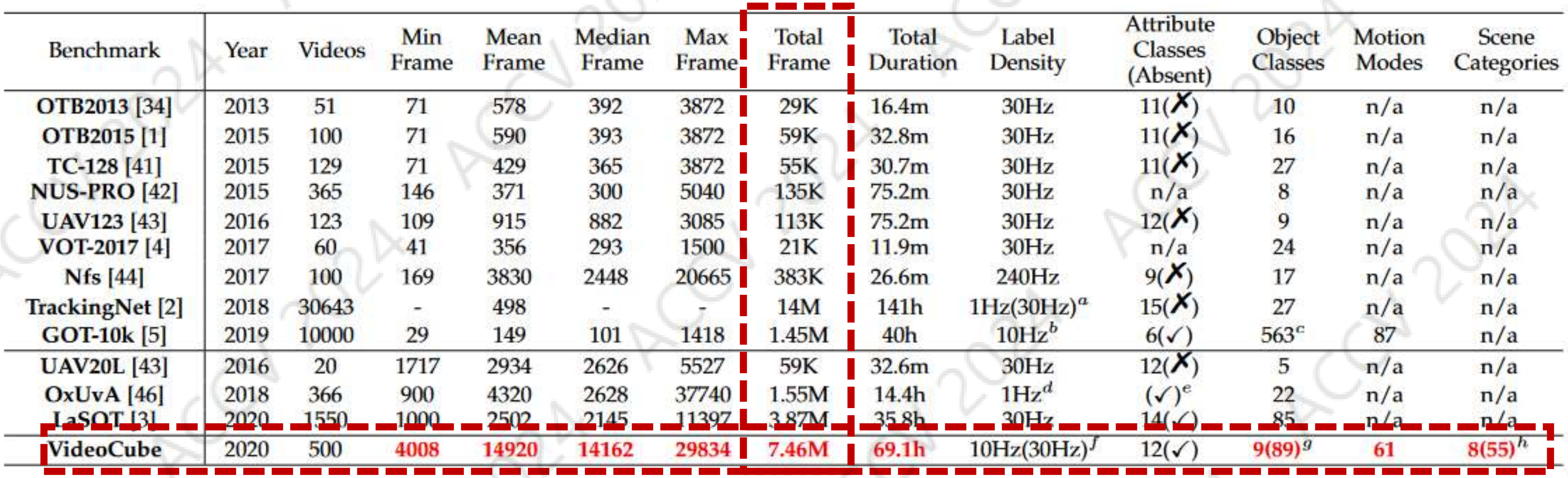
ACCV 2024

General Datasets: Large-scale GIT

- **VideoCube (2023)**

➢ **Large-scale dataset:**

- One of the largest SOT dataset currently, with an overall size **2~200 times** that of existing datasets



Benchmark	Year	Videos	Min Frame	Mean Frame	Median Frame	Max Frame	Total Frame	Total Duration	Label Density	Attribute Classes (Absent)	Object Classes	Motion Modes	Scene Categories
OTB2013 [34]	2013	51	71	578	392	3872	29K	16.4m	30Hz	11(✗)	10	n/a	n/a
OTB2015 [1]	2015	100	71	590	393	3872	59K	32.8m	30Hz	11(✗)	16	n/a	n/a
TC-128 [41]	2015	129	71	429	365	3872	55K	30.7m	30Hz	11(✗)	27	n/a	n/a
NUS-PRO [42]	2015	365	146	371	300	5040	135K	75.2m	30Hz	n/a	8	n/a	n/a
UAV123 [43]	2016	123	109	915	882	3085	113K	75.2m	30Hz	12(✗)	9	n/a	n/a
VOT-2017 [4]	2017	60	41	356	293	1500	21K	11.9m	30Hz	n/a	24	n/a	n/a
Nfs [44]	2017	100	169	3830	2448	20665	383K	26.6m	240Hz	9(✗)	17	n/a	n/a
TrackingNet [2]	2018	30643	-	498	-	-	14M	141h	1Hz(30Hz)^{a}	15(✗)	27	n/a	n/a
GOT-10k [5]	2019	10000	29	149	101	1418	1.45M	40h	10Hz^{b}	6(✓)	563^{c}	87	n/a
UAV20L [43]	2016	20	1717	2934	2626	5527	59K	32.6m	30Hz	12(✗)	5	n/a	n/a
OxUvA [46]	2018	366	900	4320	2628	37740	1.55M	14.4h	1Hz^{d}	(✓)^{e}	22	n/a	n/a
LaSOT [3]	2020	1550	1000	2502	2145	11397	3.87M	35.8h	30Hz	14(✓)	85	n/a	n/a
VideoCube	2020	500	**4008**	**14920**	**14162**	**29834**	**7.46M**	**69.1h**	10Hz(30Hz)^{f}	12(✓)	**9(89)**^{g}	**61**	**8(55)**^{h}

Comparison of VideoCube and other representative SOT benchmarks in various statistical dimensions

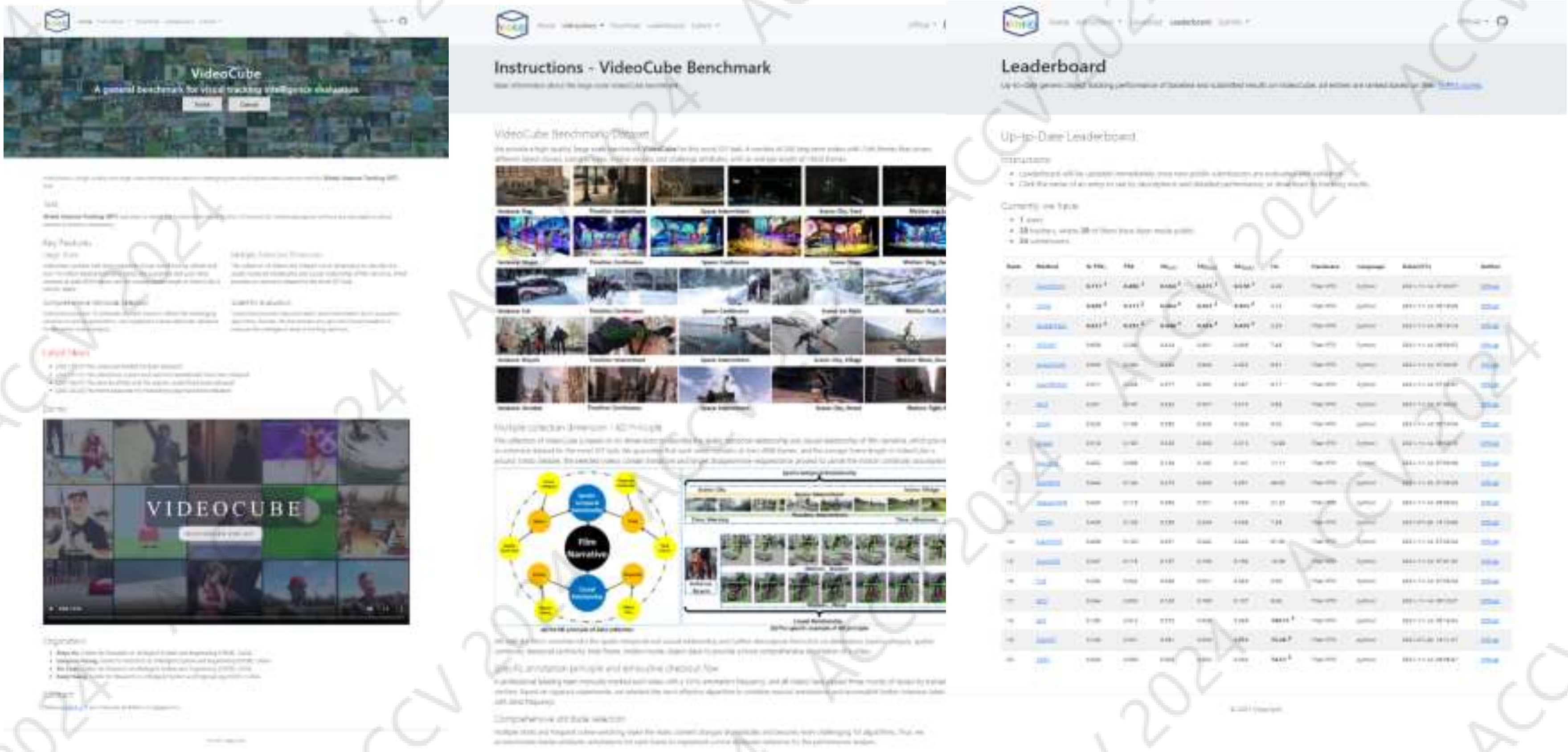
S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.

ACCV 2024

General Datasets: Large-scale GIT

- **VideoCube (2023)**

➢ Complete evaluation platform, real-time rankings, open-source toolkit.


VideoCube
A general benchmark for visual tracking intelligence evaluation
VIDEOCUBE
Instructions - VideoCube Benchmark
Leaderboard
Up-to-Date Leaderboard
Film Narrative

http://videocube.aitestunion.com/

S. Hu, X. Zhao#, *L. Huang, et al., “Global instance tracking: Locating target more like humans,” IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

ACCV 2024

Specialized Datasets

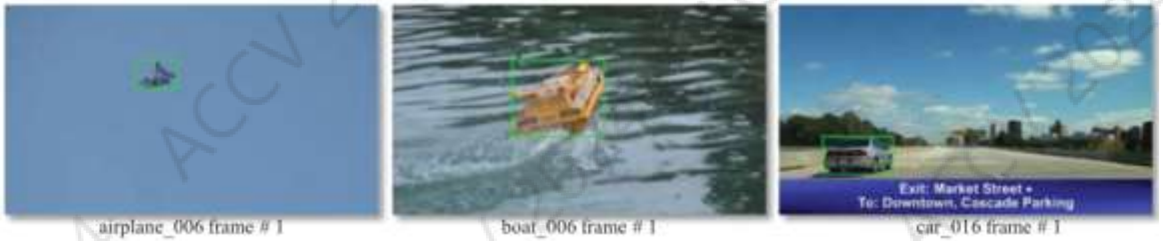
Specialized datasets are designed with specific tracking challenges in mind, focusing on unique scenarios or target types.

ACCV 2024

Specialized Datasets: Specific Object

Specialized Evaluation Environments:
Specialized environments are designed for specific tasks or special target types and are characterized by their focus on “**small but precise**” datasets. These environments aim to measure tracking performance under specific evaluation requirements and unique scenarios.

- ## NUS-PRO (2016)



airplane_006 frame # 1 boat_006 frame # 1 car_016 frame # 1

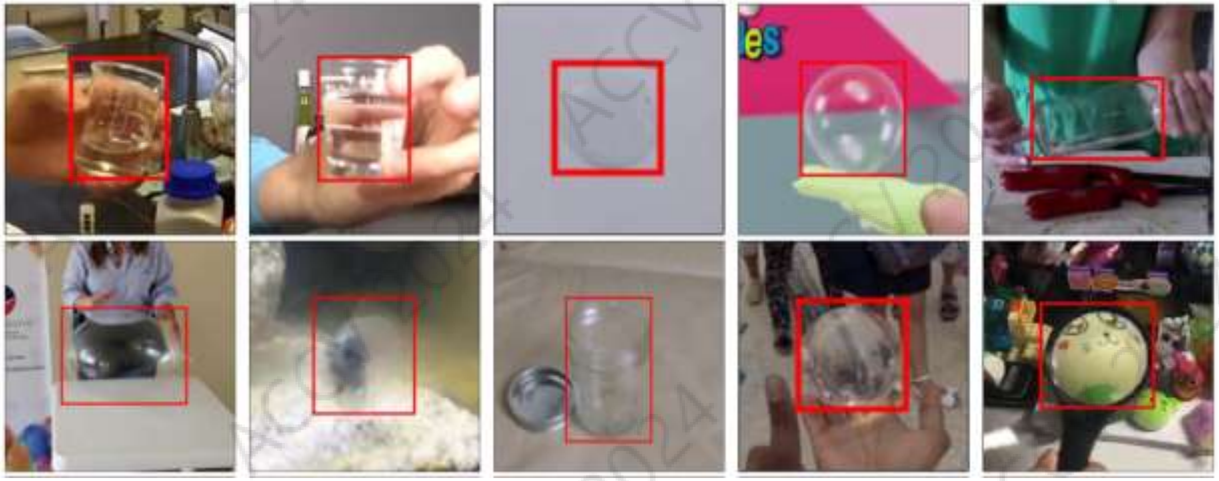
➢ NUS-PRO was developed to test tracking algorithms with a focus on two major target types: **people and rigid objects**, providing challenging sequences for tracking under occlusion conditions.

  - **Improved Understanding of Occlusion:** NUS-PRO is particularly useful for evaluating the performance of algorithms in handling occlusion, a common challenge in real-world tracking scenarios.
  - **Benchmark for Rigid and Non-rigid Tracking:** By including both people (non-rigid) and rigid objects, the dataset provides a versatile platform for testing the adaptability of tracking algorithms across different target types.

Li A, Lin M, Wu Y, et al. Nus-pro: A new visual tracking challenge[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 335-349.

Specialized Datasets: Specific Object

- **TOTB (2021)**



➢ TOTB is designed specifically for tracking transparent objects, which pose significant challenges due to their weak appearance information and **sensitivity to background interference**.

  - **Challenging Scenarios: 67.5% of the video sequences in TOTB contain background clutter**, further complicating the tracking task and emphasizing the robustness required by algorithms to handle these scenarios.
  - **Evaluation of Algorithms in Complex Scenarios:** TOTB challenges current tracking algorithms to handle transparency and complex backgrounds, providing a benchmark to **test robustness in more difficult conditions**.

Fan H, Miththanthaya H A, Rajan S R, et al. Transparent object tracking benchmark[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10734-10743.

Specialized Datasets: Specific Scenario

- **UAV123 & UAV20L (2016)**

- UAV123 was created to evaluate the performance of object tracking algorithms in **challenging aerial scenarios captured from UAVs**. UAV20L is a subset of the UAV123 dataset, but focuses on long-duration video sequences to evaluate how well tracking algorithms can handle extended tracking sessions without losing the target.
  - **Aerial Perspective:** The dataset emphasizes tracking from an aerial perspective, which introduces challenges like **object size reduction** and **frequent occlusions** due to camera movement.
  - **Emphasis on Real-time Processing:** The challenging sequences require fast and efficient algorithms, pushing the boundaries of **real-time object tracking**.

Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 445-461.

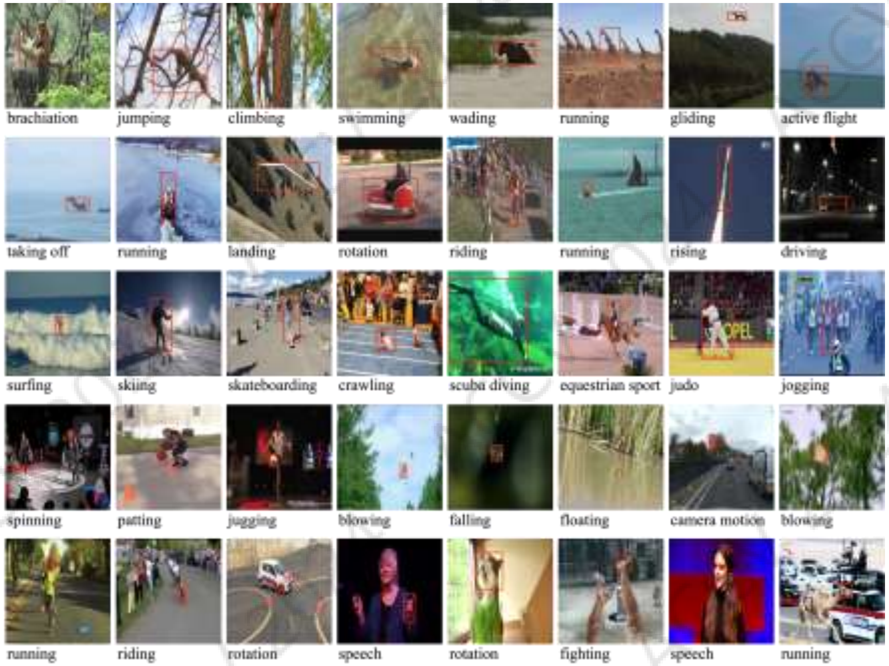
ACCV 2024

Specialized Datasets: Specific Scenario

- **BioDrone (2024)**

➢ **Motivation:** Limitations of the existing experimental environment in terms of robustness:

- Mainly focus on general scenarios, **ignoring the attention of highly challenging special scenarios**
- Mainly based on fixed lenses or handheld lenses to record moving targets, resulting in a short distance between the lens and the target, **lack of small targets and fast motion challenges**


brachiation
jumping
climbing
swimming
wading
running
gliding
active flight
taking off
running
landing
rotation
riding
running
rising
driving
surfing
skiing
skateboarding
crawling
scuba diving
equestrian sport
judo
jogging
spinning
patting
juggling
blowing
falling
floating
camera motion
blowing
running
riding
rotation
speech
rotation
fighting
speech
running

*X. Zhao, S. Hu#, Y. Wang, et al., "Biodrone: A bionic drone-based single object tracking benchmark for robust vision," International Journal of Computer Vision (**IJCV**), 2024.*

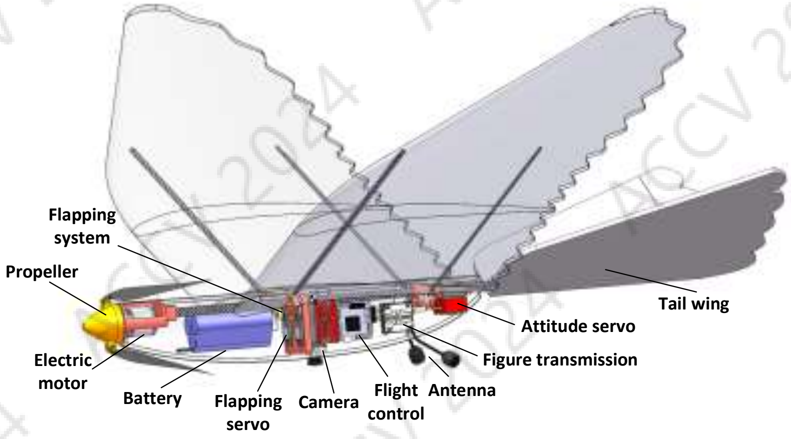
ACCV 2024

Specialized Datasets: Specific Scenario

- **BioDrone (2024)**

➢ **Robust Vision Research Dataset:**

  - The first SOT dataset from the perspective of a **bionic flapping-wing drone.**
  - The aerodynamic structure of a bionic flapping-wing drone is different from that of a traditional fixed-wing or rotary-wing drone, and there is **severe jitter between the shots.**


Flapping system
Propeller
Electric motor
Battery
Flapping servo
Camera
Flight control
Antenna
Figure transmission
Attitude servo
Tail wing

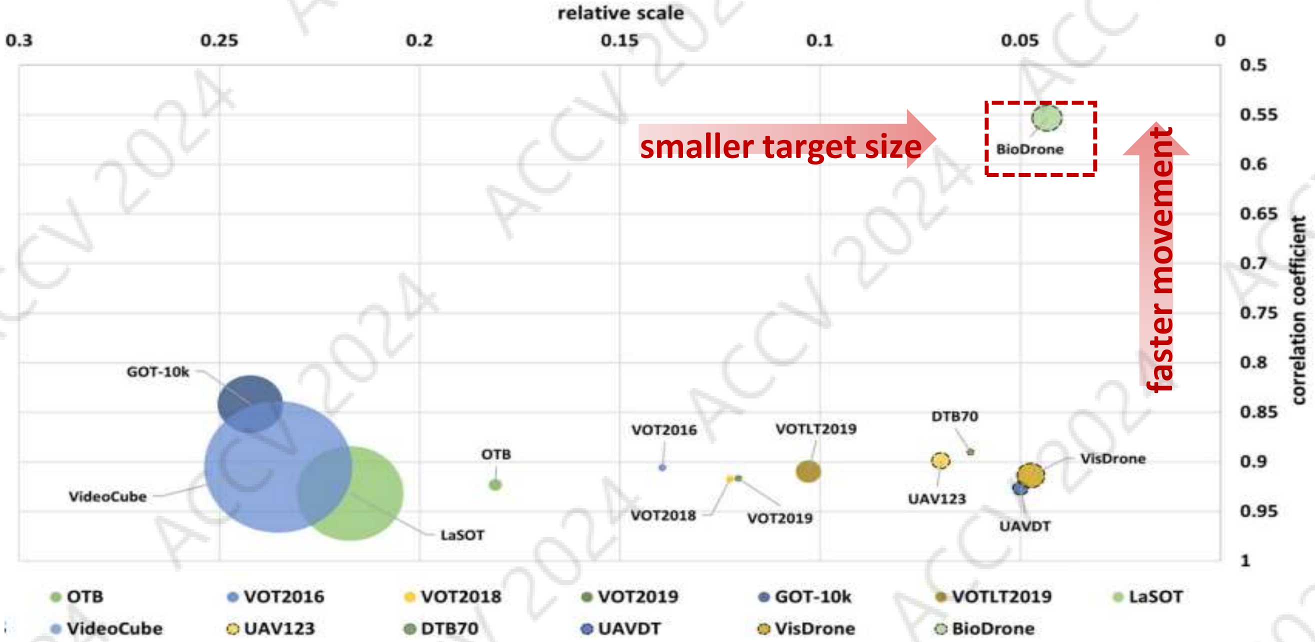
X. Zhao, S. Hu#, Y. Wang, et al., "Biodrone: A bionic drone-based single object tracking benchmark for robust vision," International Journal of Computer Vision (IJCV), 2024.

Specialized Datasets: Specific Scenario

- **BioDrone (2024)**

➢ **Robust Vision Research Dataset:**

  - Includes different flight altitudes, flight angles and flight environments, highlighting the challenges of **fast movement and small targets.**



smaller target size

faster movement

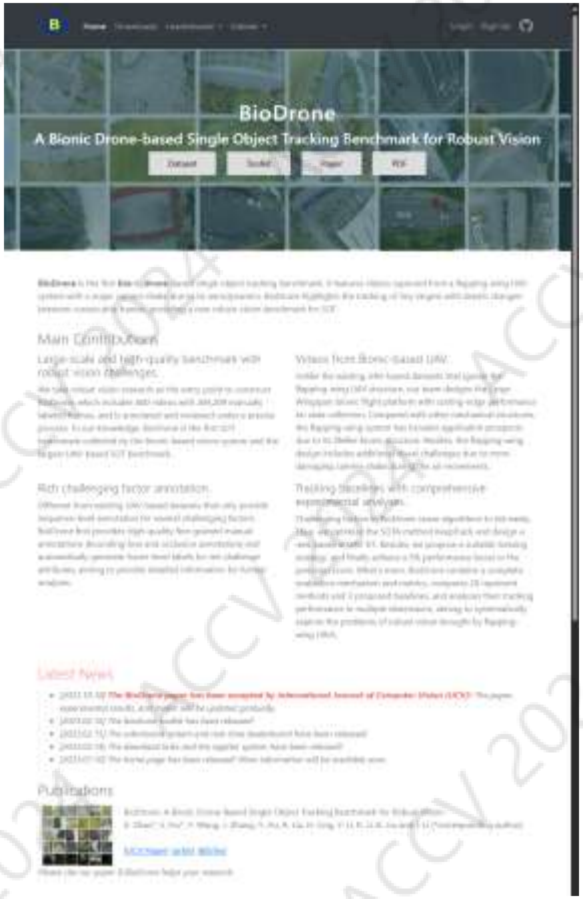
*X. Zhao, S. Hu#, Y. Wang, et al., "Biodrone: A bionic drone-based single object tracking benchmark for robust vision," International Journal of Computer Vision (**IJCV**), 2024.*

ACCV 2024

Specialized Datasets: Specific Scenario

- **BioDrone (2024)**

➢ Complete evaluation platform, real-time rankings, open-source toolkit



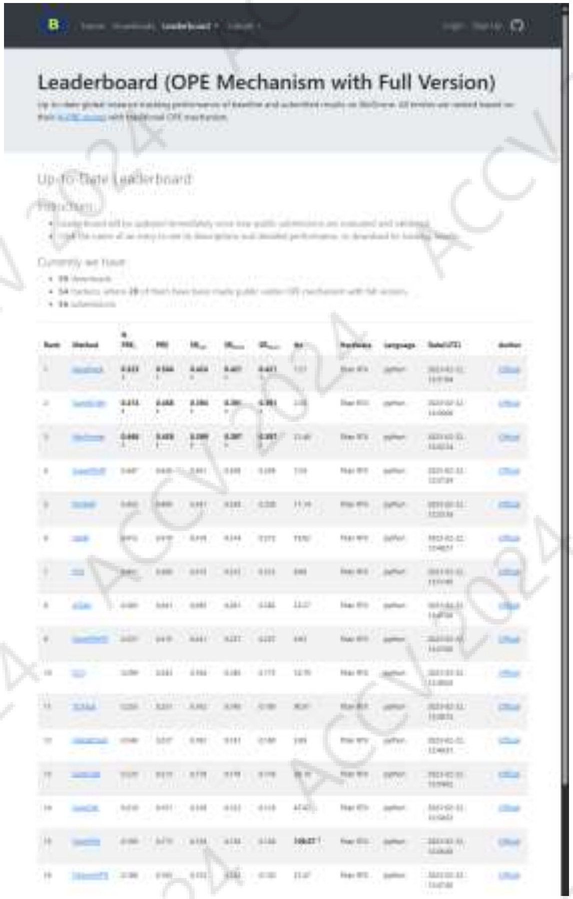
BioDrone

A Bionic Drone-based Single Object Tracking Benchmark for Robust Vision

Main Contributions

Latest News

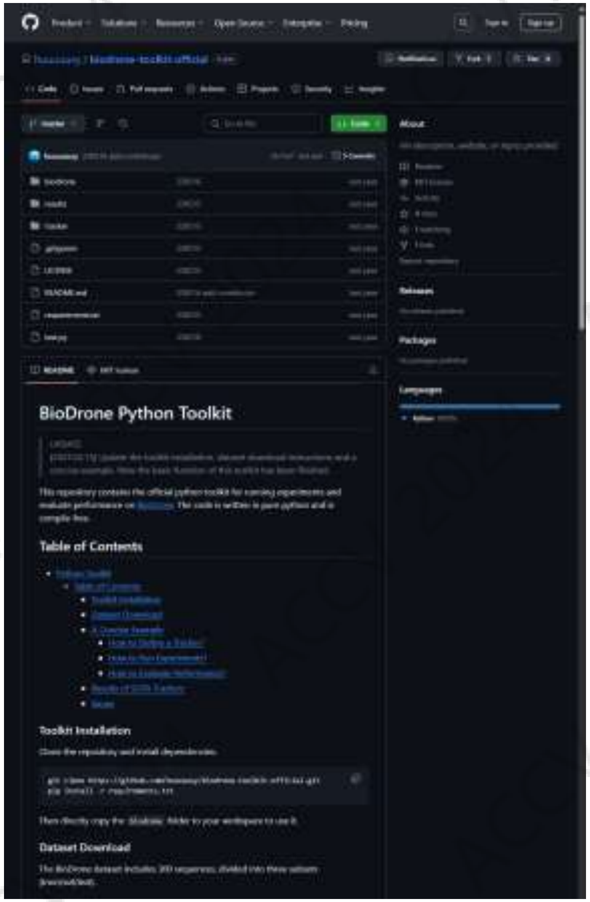
Publications



Leaderboard (OPE Mechanism with Full Version)

Up-to-Date Leaderboard

Currently we have



BioDrone Python Toolkit

Table of Contents

Toolkit Installation

Dataset Download

http://biodrone.aitestunion.com/

*X. Zhao, S. Hu#, Y. Wang, et al., "Biodrone: A bionic drone-based single object tracking benchmark for robust vision," International Journal of Computer Vision (**IJCV**), 2024.*

ACCV 2024

Competition Datasets

Competition datasets provide standardized benchmarks for comparing the performance of tracking algorithms under controlled and real-world conditions.

ACCV 2024

Competition Datasets

- **VOT-ST -> VOT-LT -> VOT-RGBT / VOT-RGBD**

VOT 10
visual object tracking

To commemorate 10 years of VOT challenges, the VOT Innitiative has set up a short online exhibition.

VOT (Visual Object Tracking) Challenge:
The VOT Challenge is an annual event established in 2013. It is **one of the most influential competitions** in the field of visual object tracking, providing standardized evaluation datasets and protocols.

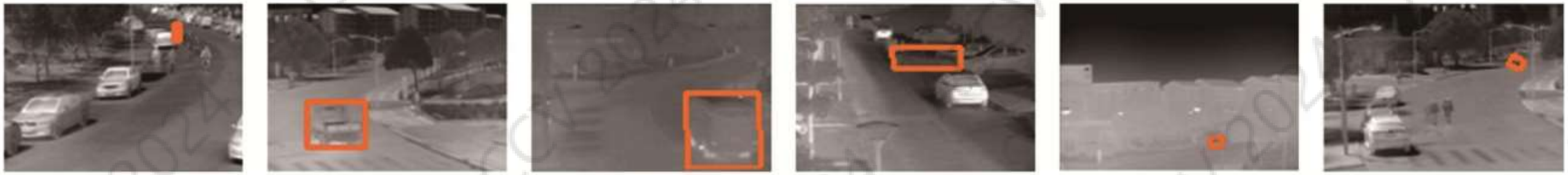
➢ **VOT-ST competition:** employs **rotated bounding boxes** or **segmentation**, supporting research in joint target segmentation and tracking tasks.

ACCV 2024

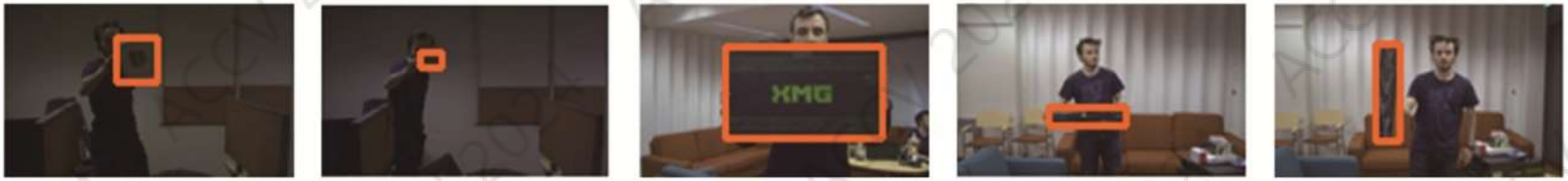
Competition Datasets

- VOT-ST -> VOT-LT -> VOT-RGBT / VOT-RGBD

➢ **VOT-LT competition:** allowing target disappearance as the distinguishing criterion between short-term and long-term tracking, and collected **50 long video sequences** as competition data.



➢ **VOT-TIR and VOT-RGBT competitions:** based on **thermal imaging** to conduct target tracking. Thermal imaging information is **less affected by lighting**, so it can still provide environmental information under **special lighting conditions.**



➢ **VOT-D and VOT-RGBD competitions:** focus on **depth information**, which can effectively separate foreground and background while providing additional support for **target occlusion issues.**

https://www.votchallenge.net/

Conclusion

➢ **General Evaluation Environments:**

- General evaluation environments feature an early start in research, numerous representative works, and wide data coverage.
- These environments aim to **provide a comprehensive experimental platform** for evaluating the **overall capabilities** of tracking algorithms in **general scenarios**.

➢ **Specialized Evaluation Environments:**

- Specialized environments are designed for **specific tasks or special target types** and are characterized by their focus on “**small but precise**” datasets.
- These environments aim to measure tracking performance under **specific evaluation requirements and unique scenarios**.

ACCV 2024

Conclusion

➢ **Competition-Based Evaluation Environments:**

- Competition environments are released as part of tracking competitions.
- These environments usually feature **highly challenging** video sequences designed to **rapidly expose algorithm weaknesses**.
- The goal is to rank participating algorithms based on **multiple performance dimensions.**

The goal is to help researchers understand the characteristics and focus of each environment, enabling them to build evaluation settings that are better suited for the specific evaluation objectives.

ACCV 2024

CONTENTS

- Introduction
- Part 1. Visual Object Tracking Task
- Part 2. Experimental Environments
- **Part 3. Algorithms and Traditional Machine-Machine Comparisons**
  - **Machines**
  - **Machine-to-Machine Evaluation**
- Part 4. Human Visual Abilities and Visual Turing Test
- Trends and Future Directions

Scan to download this tutorial PPT

ACCV 2024

Machines

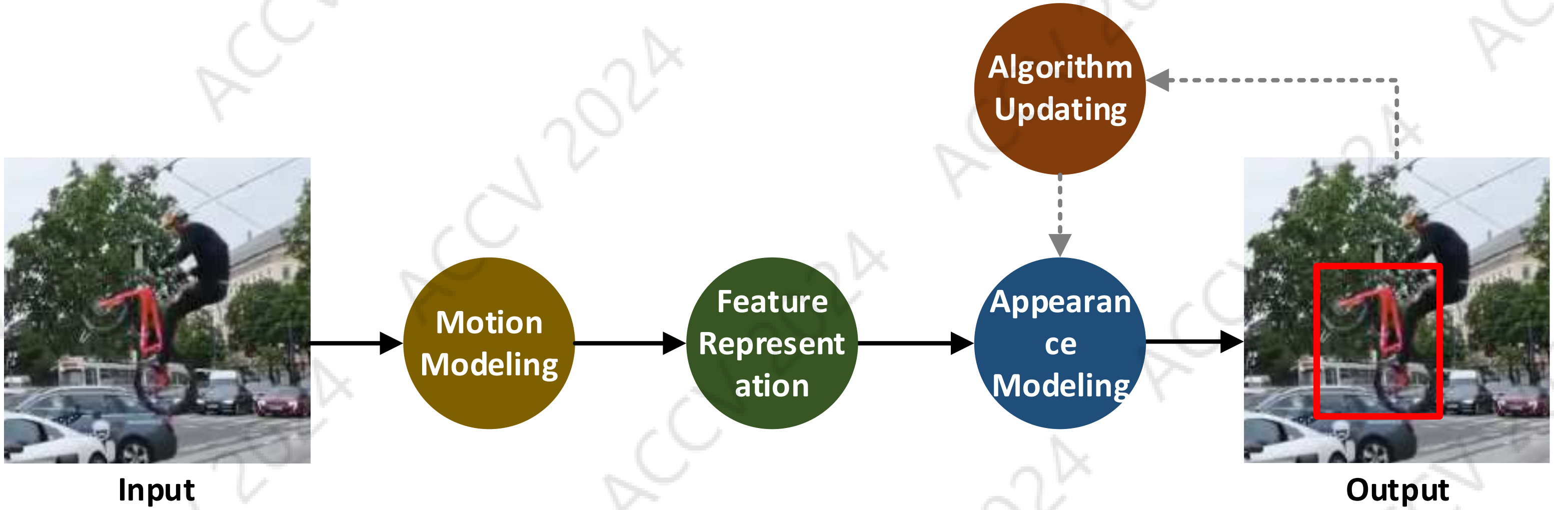
We will explore different types of machine-based tracking systems, including traditional algorithms and more advanced deep learning models.

ACCV 2024

Machines: Traditional Trackers

Traditional trackers :

Traditional trackers usually includes the following steps: **motion modeling, feature representation, appearance modeling, and algorithm updating**.



Algorithm Updating

Motion Modeling

Feature Representation

Appearance Modeling

Input

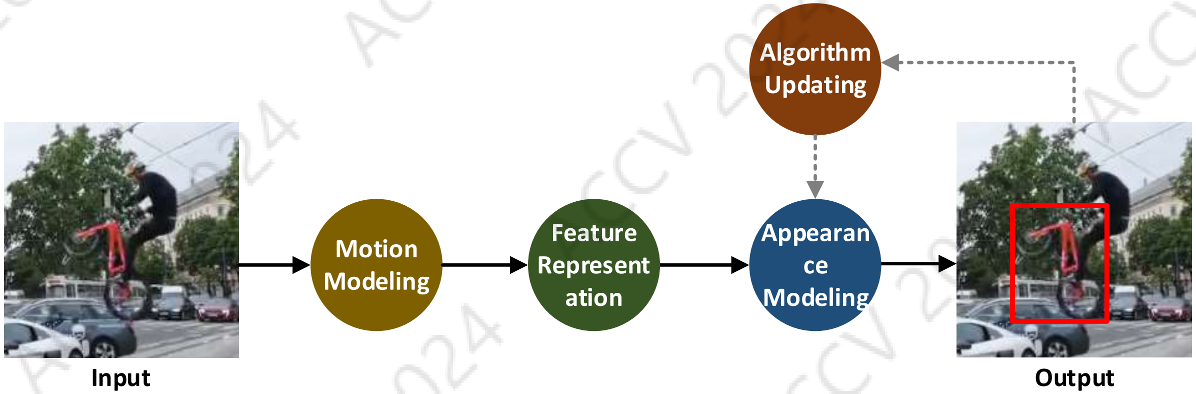
Output

➢ Motion modeling

- **Purpose:** Predict the target trajectory in subsequent frames by estimating the position state of the target.
- **Representative Methods:** Particle filtering, Sliding window

ACCV 2024

Machines: Traditional Trackers



Input

Motion Modeling

Feature Representation

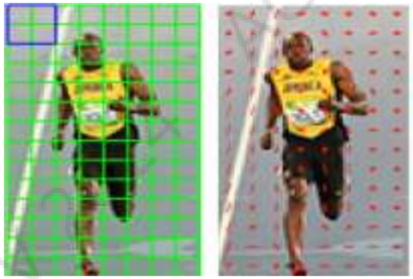
Appearance Modeling

Algorithm Updating

Output

- **Feature Representation**
  - **Global Features:** Early methods extracted features from the **entire target**.
    - ☐ Grayscale features, gradient histogram features, and color histogram features

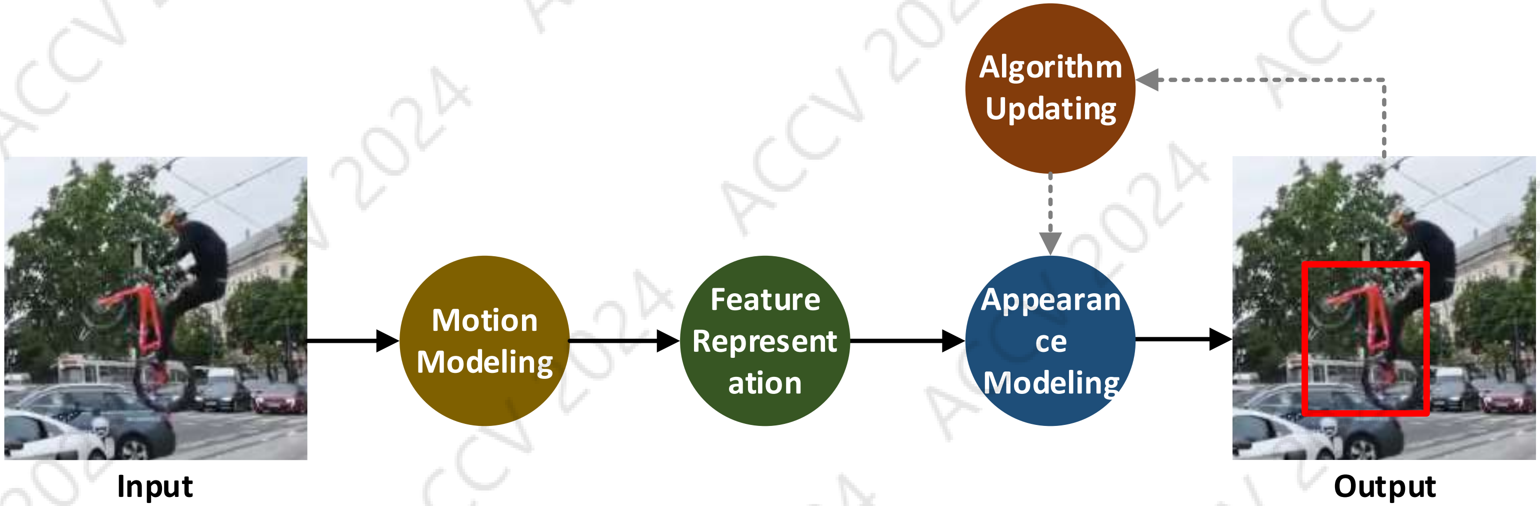
grayscale features



gradient histogram features

  - **Local Features:** To **handle challenges** like occlusion and deformation, local feature extraction methods were applied.
    - ☐ Segment the target into independent regions and fuse information from each part

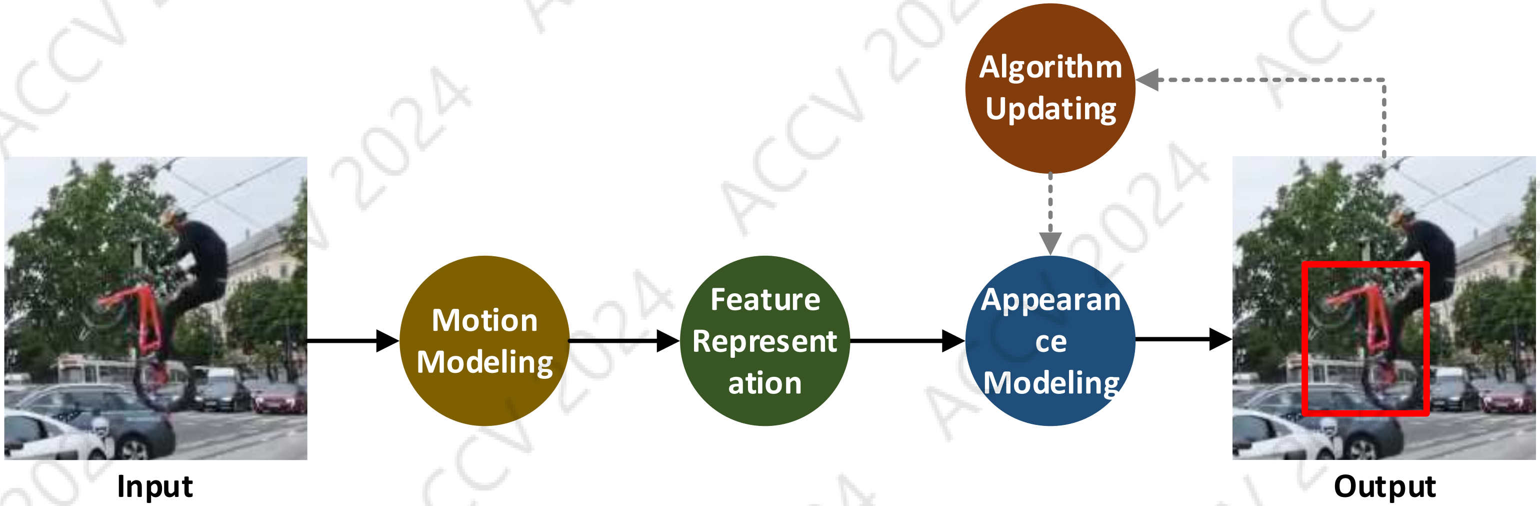
Machines: Traditional Trackers


Algorithm Updating
Input
Motion Modeling
Feature Representation
Appearance Modeling
Output

- **Appearance Modeling**
  - **Generative Methods:**
    - ☐ First, maintain a target template set using methods like incremental subspace or block sparse representation.
    - ☐ Then, measure **similarity** based on the **distance** between the candidate sample and the target template set.
  - **Discriminative Methods:**
    - ☐ Treat tracking as a **classification problem**, classifying between the target and the background.

ACCV 2024

Machines: Traditional Trackers



Algorithm Updating

Motion Modeling

Feature Representation

Appearance Modeling

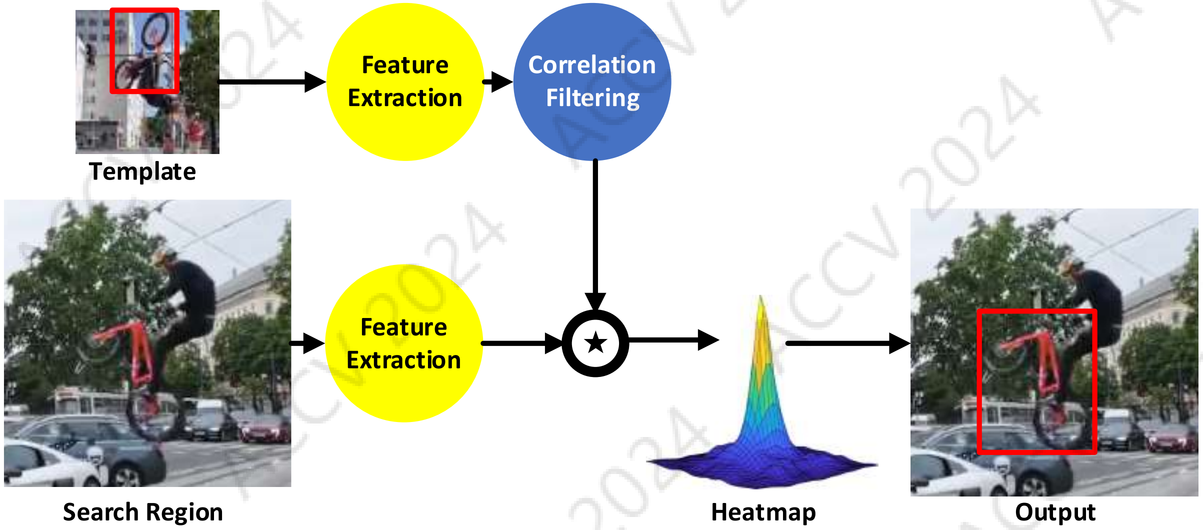
Input

Output

- **Algorithm Updating**
  - **Purpose:**
    - ☐ The initial static template struggles to **continuously guide tracking for dynamically changing targets**.
    - ☐ The update strategy ensures that the algorithm **adapts to changes** in the target's appearance.
  - **Representative Methods:**
    - ☐ Incremental learning-based updates
    - ☐ Online refactoring of the appearance model

ACCV 2024

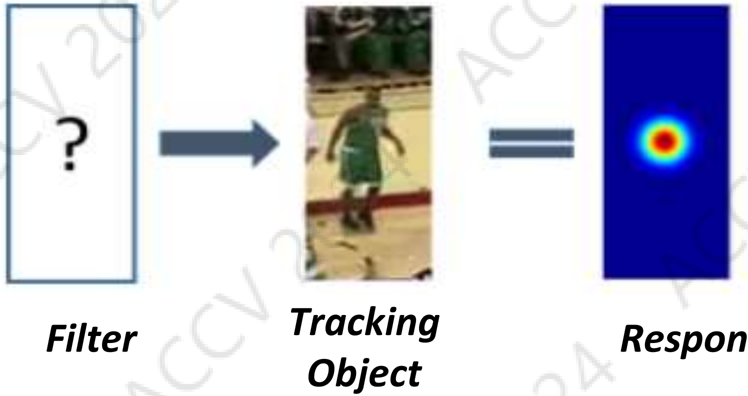
Machines: Correlation Filter Based Trackers


Feature Extraction
Correlation Filtering
Template
Feature Extraction
★
Search Region
Heatmap
Output

- **Overall Process:** Feature extraction, correlation filtering, output prediction
- **Advantages:**
  - **Correlation filter theory** expands training samples through cyclic shifts, effectively solving the problem of insufficient data in early methods.
  - **Fast Fourier Transform** reduces computational load and improves tracking efficiency.
- **Representative Methods：** KCF, ECO, UPDT

ACCV 2024

Machines: Correlation Filter Based Trackers



- **Kernel Correlation Filter (KCF) Algorithm**
  - A typical discriminative object tracking method.
  - **Determines the target position by training a filter**, with high computational efficiency and strong tracking performance.
  - **Core Design:**
    - ☐ **Target Initialization (t frame)**: Select the target and sample around it to train the classifier (filter).
    - ☐ **Target Position Update (t+1 frame)**: Sample near the target in the t+1 frame, use the classifier to perform correlation operations, and calculate the response at the sampling points.
    - ☐ **Determine Target Position**: Identify the sampling point with the **strongest response** that meets the threshold, and treat it as the target position in the t+1 frame.

Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(3): 583-596..

ACCV 2024

Machines: Correlation Filter Based Trackers

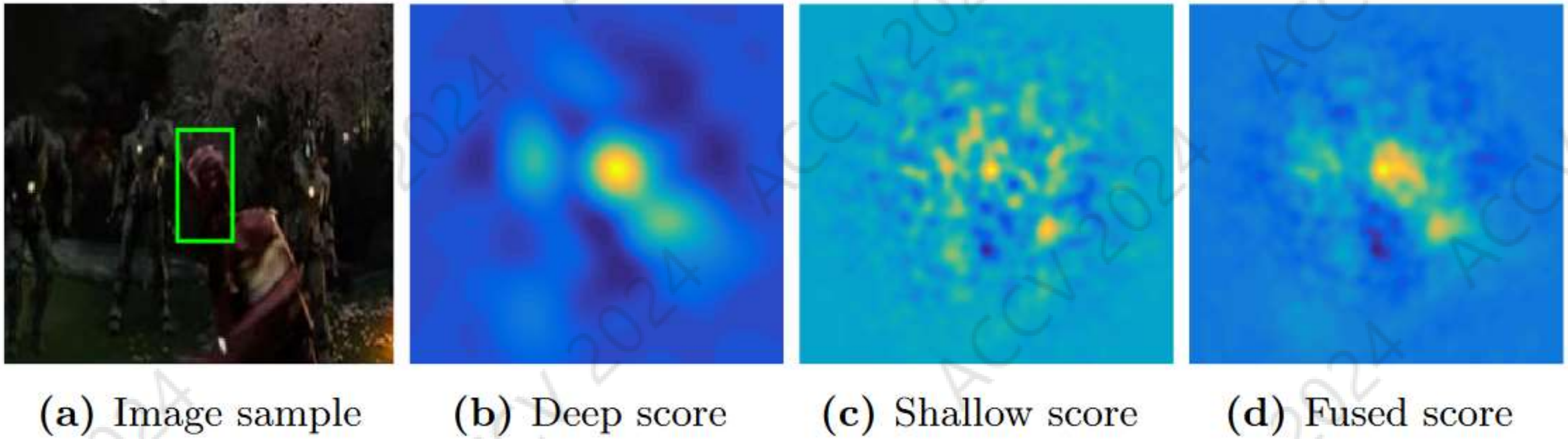
Our Representation
Component 1
Component 2
Component 3
Component 4
Baseline

- **Efficient Convolutional Operators (ECO) Algorithm**
  - Aims to address the **computational complexity and overfitting issues** in Discriminative Correlation Filter (DCF) methods.
  - **Core Design:**
    - ☐ **Factorized Convolution Operators:** Reduced model parameters, lowering complexity and avoiding overfitting.
    - ☐ **Generative Sample Space Model:** A compact sample generation model that enhances the diversity of training samples.
    - ☐ **Conservative Model Update Strategy:** By reducing the frequency of model updates, it improves tracking speed and prevents model drift.

Danelljan M, Bhat G, Shahbaz Khan F, et al. Eco: Efficient convolution operators for tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6638-6646.

ACCV 2024

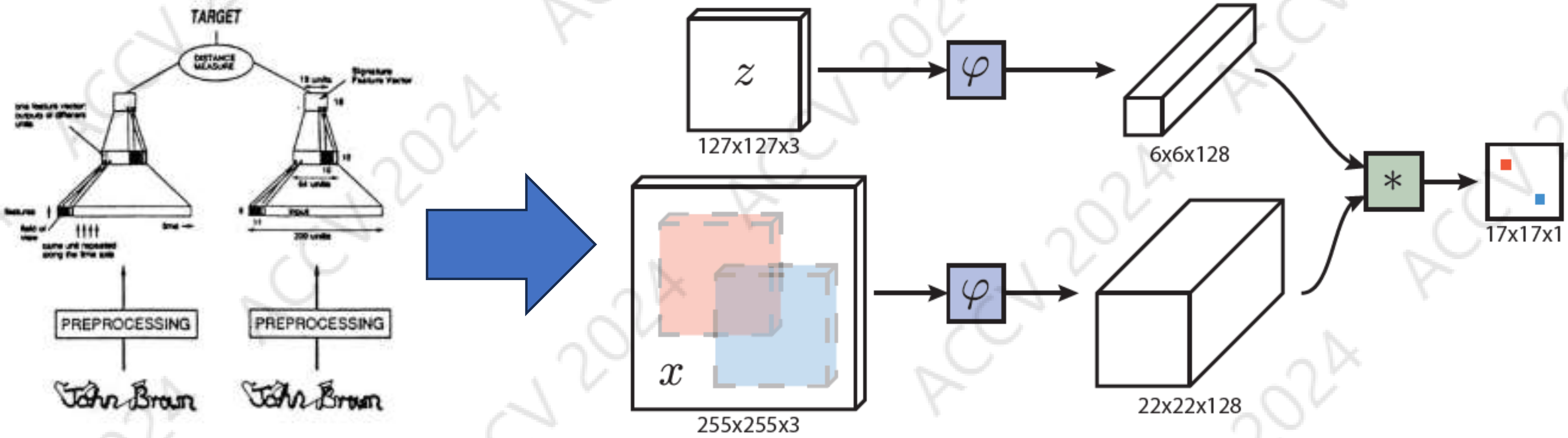
Machines: Correlation Filter Based Trackers


(a) Image sample (b) Deep score (c) Shallow score (d) Fused score

- **Unveiling the Power of Deep Tracking (UPDT) Algorithm**
  - Solves the issue in ECO where **deep features were not fully utilized**.
  - **Core Design:**
    - ☐ **Separation of Deep and Shallow Features:** Deep features model high-level semantic information, while shallow features model texture and color information.
    - ☐ **Adaptive Response Map Fusion:** Based on detection quality assessment, adaptively fuses the response maps of deep and shallow features with weighted fusion.

Bhat G, Johnander J, Danelljan M, et al. Unveiling the Power of Deep Tracking [J]. arXiv preprint arXiv:1804.06833, 2018.

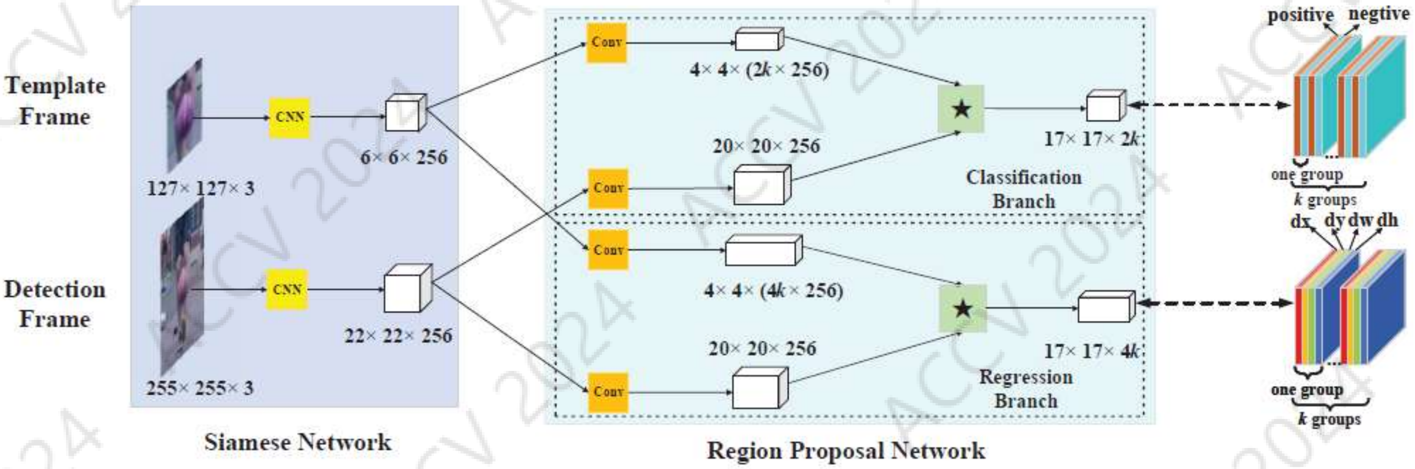
Machines: Siamese Network


TARGET
DISTANCE MEASURE
PREPROCESSING
PREPROCESSING
z
127x127x3
φ
6x6x128
$*$
17x17x1
x
255x255x3
φ
22x22x128

- **SiamFC Algorithm**
  - **Network Architecture:** SiamFC uses two identical fully convolutional networks (FCNs) to extract features from the target template in the first frame and the search region in the subsequent frames. The network computes the cross-correlation between the two feature maps to predict the location of the object in the search region.
  - **Core Design:** SiamFC assumes that the object remains within a specific search region and relies on the correlation between frames to track the object **without needing to update the model online.**
  - **Impact**: SiamFC laid the groundwork for further development in deep learning-based object tracking.

Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." Computer Vision–ECCV 2016 Workshops

Machines: Siamese Network


Template Frame
127× 127× 3
CNN
6× 6× 256
Detection Frame
255× 255× 3
CNN
22× 22× 256
Siamese Network
Conv
4× 4× (2k × 256)
Conv
20× 20× 256
★
17× 17× 2k
Classification Branch
Conv
4× 4× (4k × 256)
Conv
20× 20× 256
★
17× 17× 4k
Regression Branch
Region Proposal Network
positive
negtive
one group
k groups
dx dy dw dh
one group
k groups

SiamRPN adopts the Region Proposal Network (RPN), enabling the tracker to predict position and shape.

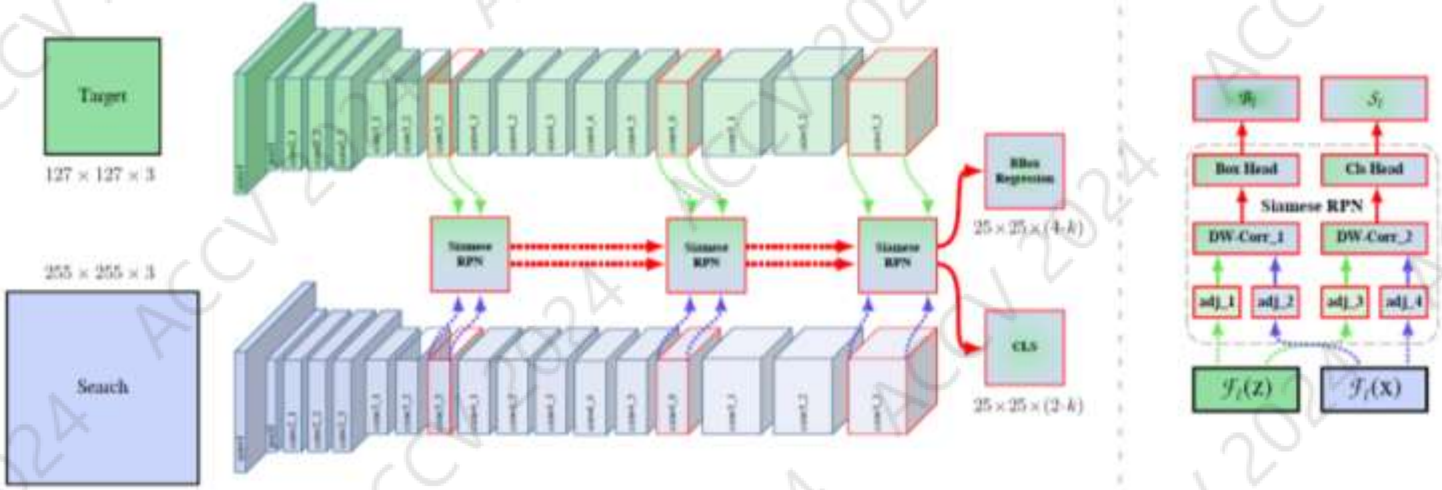
➢ SiamRPN Algorithm

- **Network Architecture:** SiamRPN consists of two fully convolutional Siamese networks to extract features from the target template and the search region. These features are then passed through the RPN, which generates **region proposals** and refines the bounding box for the tracked object.
- **Core Design:** SiamRPN treats object tracking as a **detection problem** by using RPN to predict the object's location and bounding box in each frame. This allows for more accurate localization and **bounding box regression**, improving the tracking performance, especially for objects undergoing **scale and shape changes**.

Li, Bo, et al. "High performance visual tracking with siamese region proposal network." CVPR. 2018.

ACCV 2024

Machines: Siamese Network


Target
127 × 127 × 3
255 × 255 × 3
Search
Siamese RPN
Siamese RPN
Siamese RPN
BBox Regression
25 × 25 × (4·k)
CLS
25 × 25 × (2·k)
$\mathcal{B}_l$
$\mathcal{S}_l$
Box Head
Cls Head
Siamese RPN
DW-Corr_1
DW-Corr_2
adj_1
adj_2
adj_3
adj_4
$\mathcal{F}_l(z)$
$\mathcal{F}_l(x)$

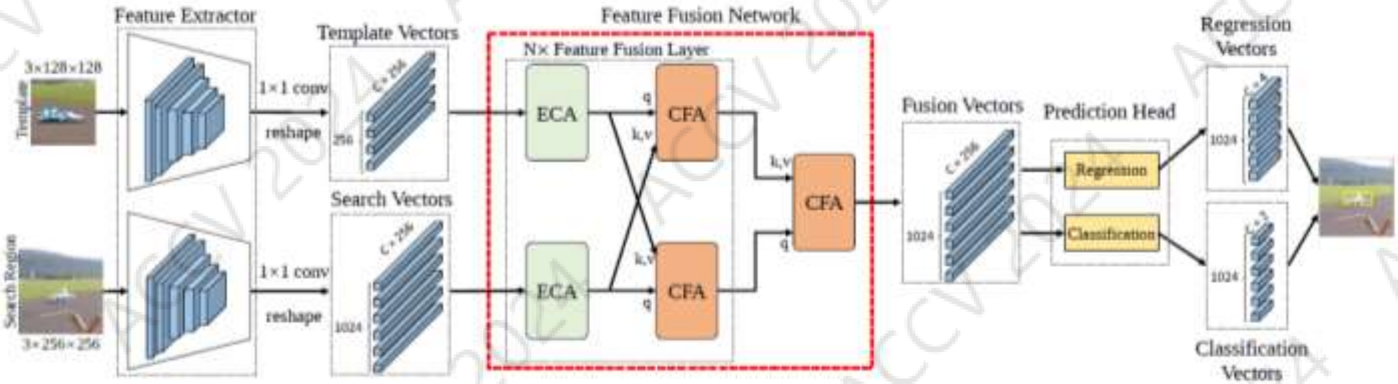
SiamRPN++ adjusted the sampling strategy, making it possible to use deep networks like ResNet-50.

➢ SiamRPN++ Algorithm

- **Network Architecture:** SiamRPN++ uses a ResNet backbone to extract more robust and discriminative features. This improves the network's ability to handle complex scenarios, including large appearance variations, occlusion, and background clutter, while maintaining real-time performance.
- **Core Design:** By upgrading the backbone from shallow convolutional networks to **deeper** ones (such as ResNet-50), SiamRPN++ can capture more detailed and **hierarchical feature representations**, which improves tracking performance in challenging conditions.

Li, Bo, et al. "Siamrpn++: Evolution of siamese visual tracking with very deep networks." CVPR. 2019.

Machines: Siamese Network

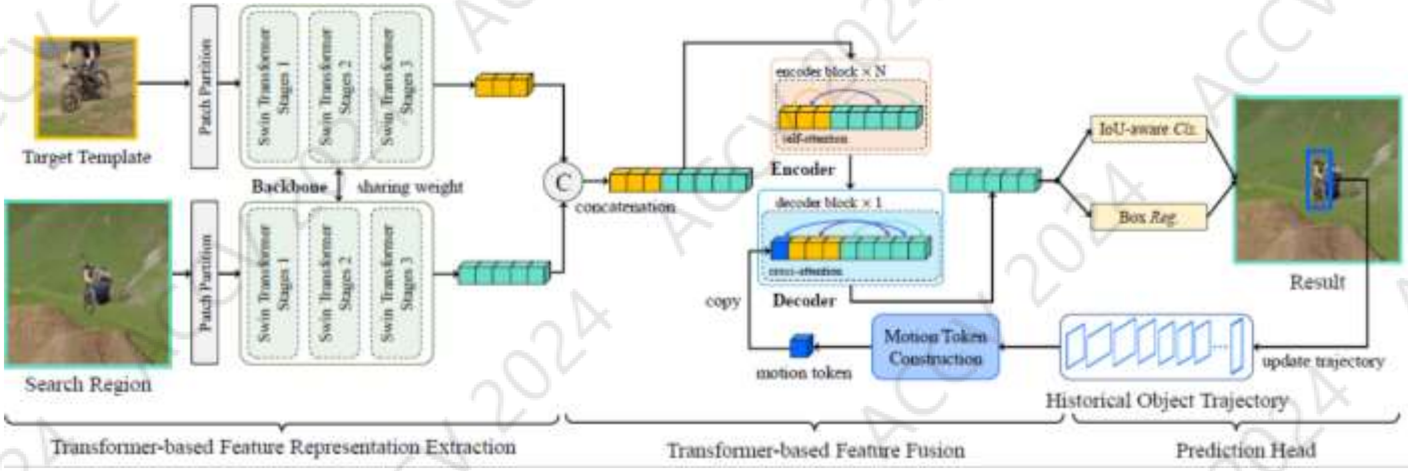

Feature Extractor
Template Vectors
Feature Fusion Network
N× Feature Fusion Layer
Regression Vectors
3×128×128
Template
1×1 conv
reshape
C=256
256
ECA
q
k,v
CFA
Fusion Vectors
Prediction Head
C=4
1024
Search Vectors
k,v
CFA
C=256
1024
Regression
Classification
Search Region
3×256×256
1×1 conv
reshape
C=256
1024
ECA
k,v
q
CFA
q
C=2
1024
Classification Vectors

TransT incorporates attention to replace the conventional correlation operation, efficiently merging the features of the template and search region.

- **TransT Algorithm**
  - **Network Architecture:** Unlike simple linear correlation in traditional trackers, TransT applies an attention mechanism in the feature fusion module to extract more comprehensive and context-aware feature representations. The **self-attention mechanism** in Transformers allows the model to capture dependencies between the target and the background more effectively.
  - **Core Design:** The central idea of TransT is to leverage Transformer networks to model **long-range dependencies** and capture **global contextual information** in the feature space, leading to more robust tracking performance, particularly in challenging scenarios like occlusions, scale variation, and background clutter.

Chen, Xin, et al. "Transformer tracking." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Machines: Siamese Network


Target Template
Patch Partition
Swin Transformer Stages 1
Swin Transformer Stages 2
Swin Transformer Stages 3
Backbone
sharing weight
Search Region
C
concatenation
encoder block × N
self-attention
Encoder
decoder block × 1
cross-attention
Decoder
copy
motion token
Motion Token Construction
Historical Object Trajectory
update trajectory
IoU-aware Cls
Box Reg
Result
Transformer-based Feature Representation Extraction
Transformer-based Feature Fusion
Prediction Head

SwinTrack adopts the Swin Transformer as its backbone, while also utilizing an attention-based feature fusion module. This represents a major leap for Transformer-based trackers.

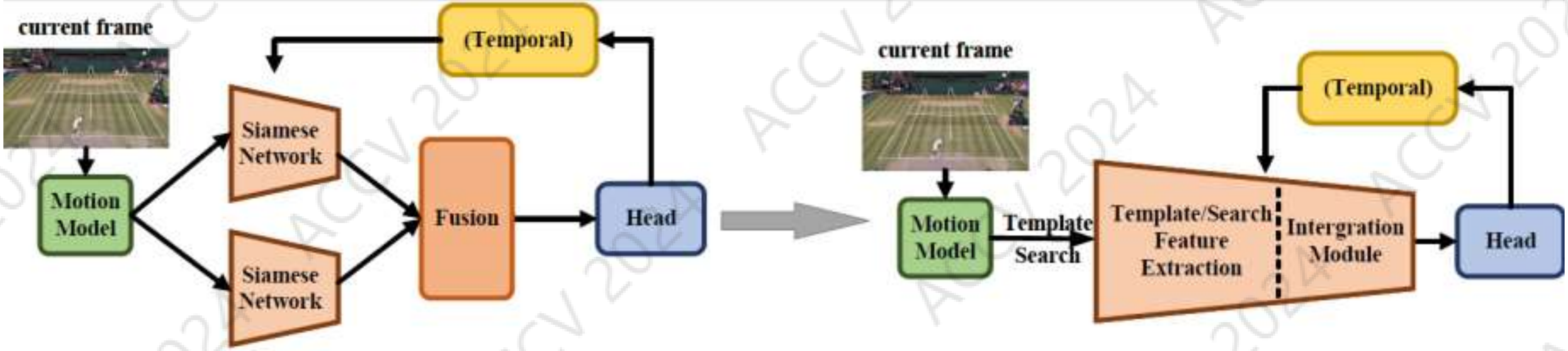
➢ **SwinTrack Algorithm**

- **Network Architecture:** SwinTrack is a state-of-the-art object tracking algorithm that leverages the **Swin Transformer** architecture for feature extraction and modeling long-range dependencies. The Swin Transformer, originally proposed for visual recognition tasks, is adapted in SwinTrack to handle the unique challenges of object tracking.
- **Core Design:** SwinTrack utilizes shifted **window-based attention** to efficiently capture both **local and global contextual information**, making it particularly effective in complex tracking scenarios involving occlusions, background clutter, and varying object scales.

Lin, Liting, et al. "Swintrack: A simple and strong baseline for transformer tracking." NeurIPS 2022.

ACCV 2024

Machines: One-Stream Trackers

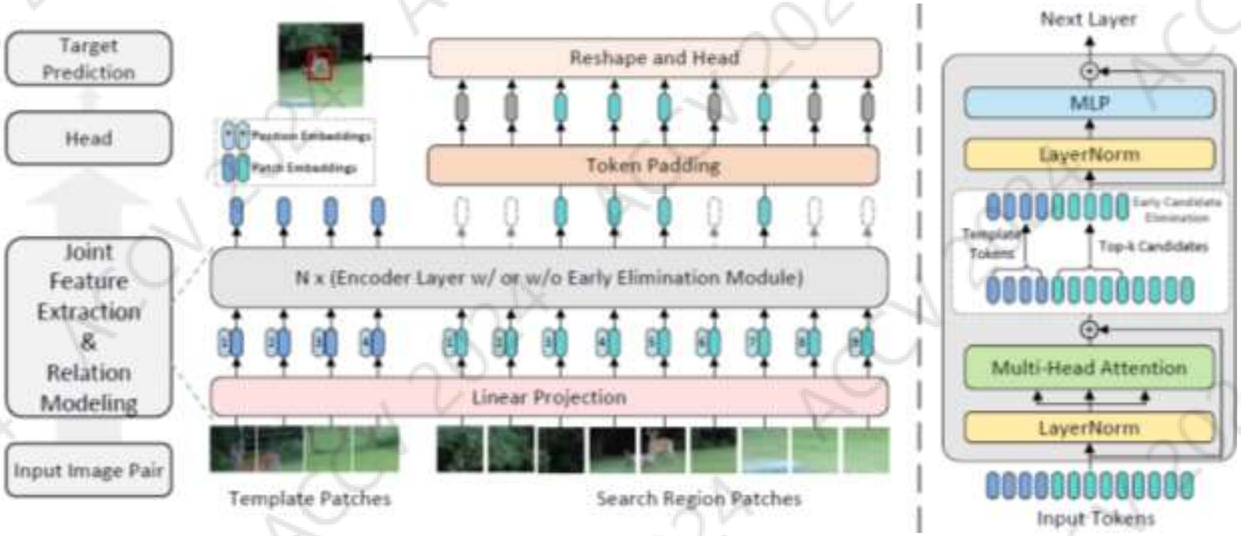

current frame
Motion Model
Siamese Network
Siamese Network
Fusion
Head
(Temporal)
current frame
Motion Model
Template
Search
Template/Search Feature Extraction
Intergration Module
(Temporal)
Head

(a) pipeline of Siamese Network (b) pipeline of One-Stream Network

- The original Siamese neural network (SiamFC) used a **CNN backbone** and a **cross-correlation layer** for feature fusion. This **two-stream approach** processes the template and search region separately, then merges the results for tracking.
- As the field of computer vision and hardware have progressed, **self-attention mechanisms** have been introduced into Siamese networks. This has led to the development of **Transformer-based one-stream architectures**, which process both the template and search region in a unified manner, ultimately replacing the Siamese structure.

ACCV 2024

Machines: One-Stream Trackers


Target Prediction
Head
Joint Feature Extraction & Relation Modeling
Input Image Pair
Reshape and Head
Position Embeddings
Patch Embeddings
Token Padding
N x (Encoder Layer w/ or w/o Early Elimination Module)
Linear Projection
Template Patches
Search Region Patches
Next Layer
MLP
LayerNorm
Early Candidate Elimination
Template Tokens
Top-k Candidates
Multi-Head Attention
LayerNorm
Input Tokens

- **OSTrack Algorithm**
  - OSTrack proposes a unified one-stream architecture where both **feature learning and relation modeling** are performed within a single network. This contrasts with traditional tracking models that may separate feature extraction from the relation modeling process, thus **reducing computational overhead and improving efficiency.**
  - **Core Design:** The framework integrates feature learning and the relationship between the target and the search region in **one stream**, ensuring that the network can simultaneously learn object representations and how they relate to the background or other objects in the scene.

Ye, Botao, et al. *"Joint feature learning and relation modeling for tracking: A one-stream framework." ECCV, 2022.*

ACCV 2024

Machines: LVM-based Trackers

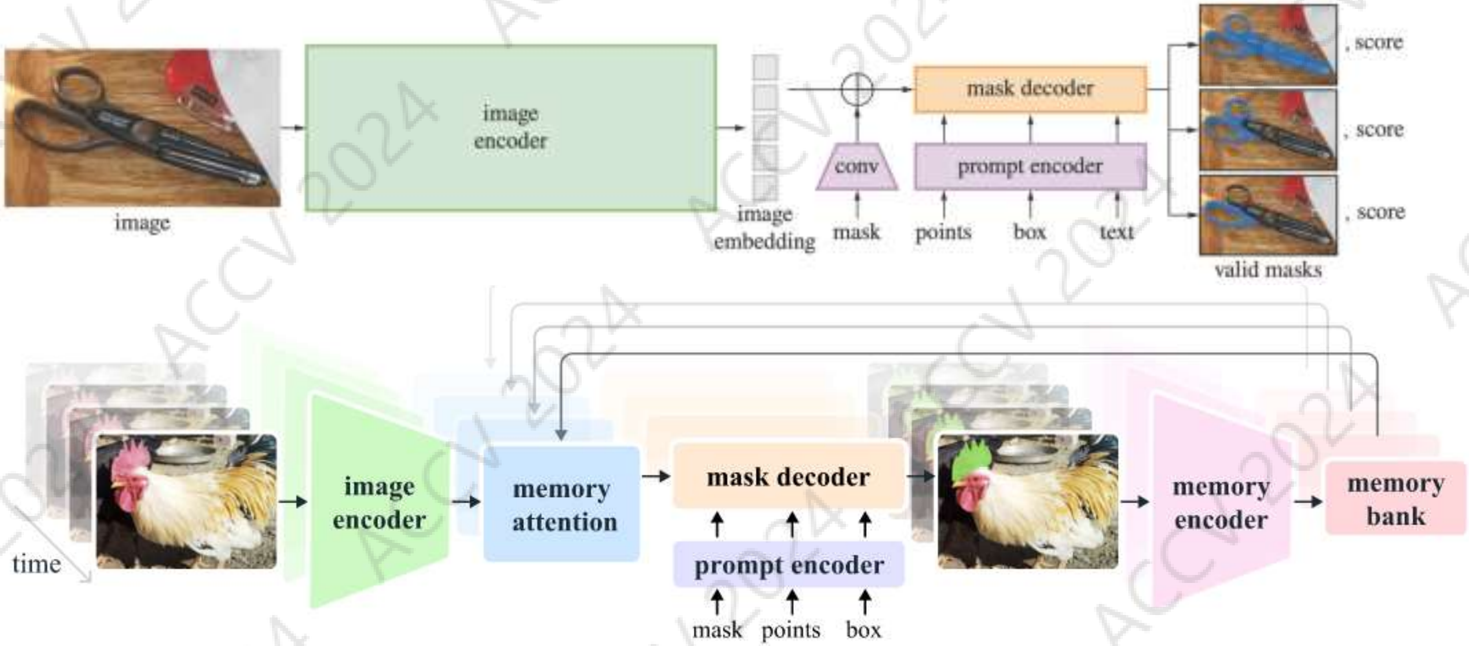


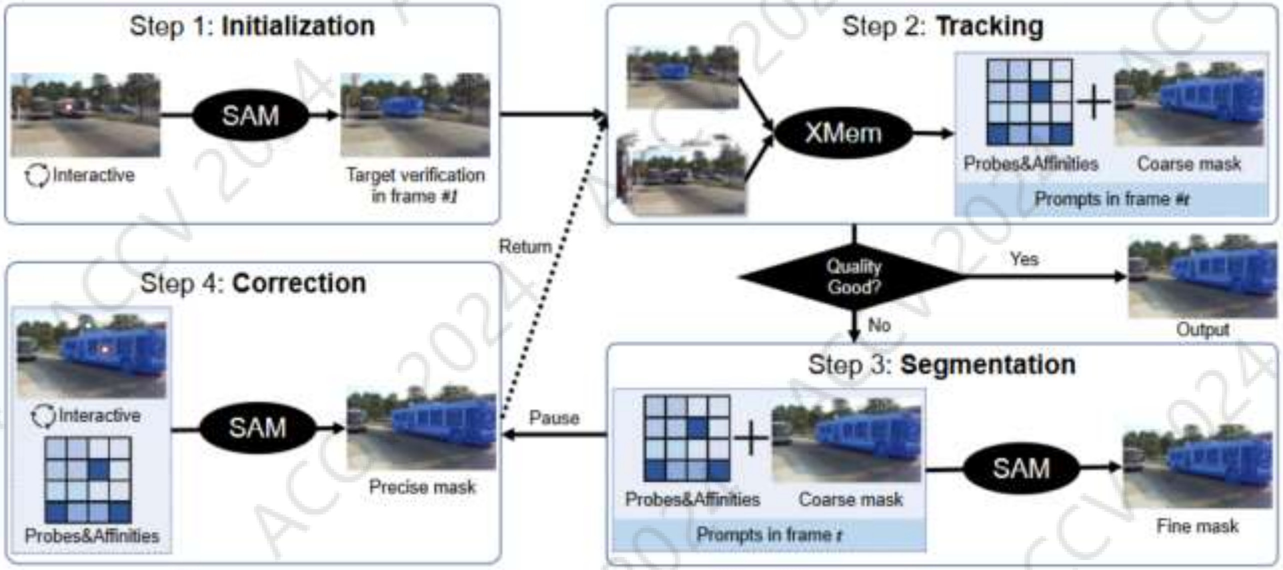
image
image encoder
image embedding
conv
mask
mask decoder
prompt encoder
points
box
text
score
score
score
valid masks
time
image encoder
memory attention
mask decoder
prompt encoder
mask
points
box
memory encoder
memory bank

SAM & SAM2

- **Background:** Recently, foundational vision/multimodal models have demonstrated exceptional capabilities in perceiving and understanding image/modality content.
  - **How can these models be applied to SOT tasks? → Design a pipeline**
- **Exploratory Works**
  - TAM (Track Anything Model)
  - SAM-Track (Segment and Track Anything)
  - TrackGPT (Tracking with Human-Intent Reasoning)

ACCV 2024

Machines: LVM-based Trackers

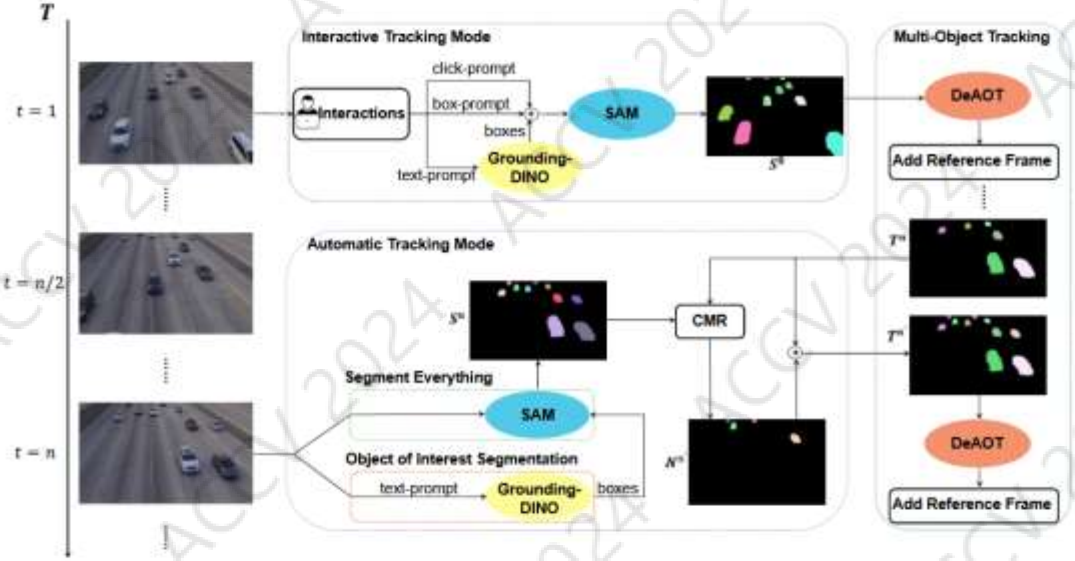

Step 1: Initialization
SAM
Interactive
Target verification in frame #1
Step 2: Tracking
XMem
Probes&Affinities
Coarse mask
Prompts in frame #t
Quality Good?
Yes
Output
No
Return
Step 4: Correction
Interactive
Probes&Affinities
SAM
Precise mask
Pause
Step 3: Segmentation
Probes&Affinities
Coarse mask
Prompts in frame t
SAM
Fine mask

- **TAM (Track Anything Model):** Combines SAM, DeAOT, and Grounding-DINO to create an efficient multi-object video segmentation model.
- **Pipeline**
  - Users **interactively initialize** by clicking on the object to define the target.
  - XMem is used to give mask predictions for the object in the next frame based on **temporal and spatial correspondence**.
  - SAM is utilized to provide a more **precise mask description**.
  - Users can **pause** and correct the tracking immediately upon noticing a failure.

Yang J, Gao M, Li Z, et al. Track anything: Segment anything meets videos[J]. arXiv preprint arXiv:2304.11968, 2023.

ACCV 2024

Machines: LVM-based Trackers

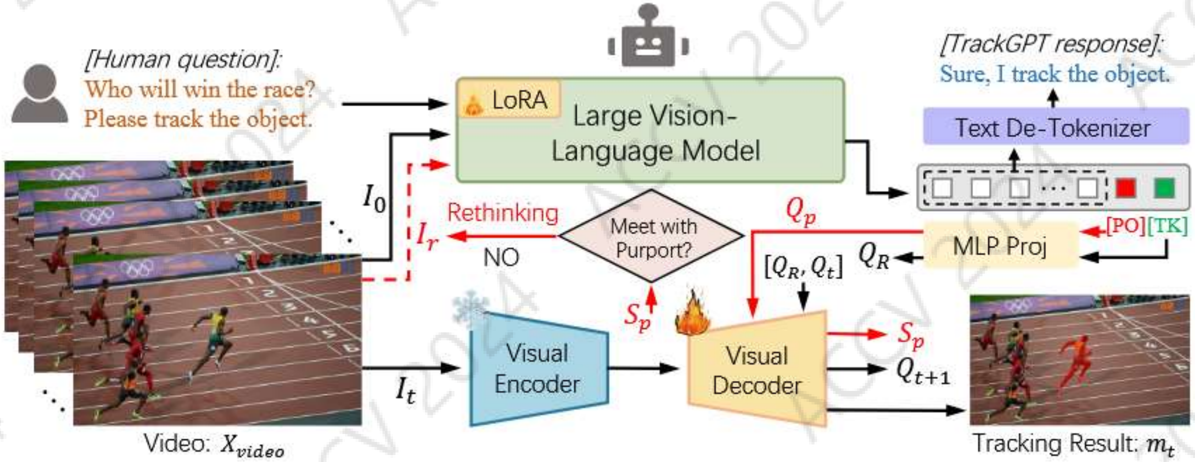

T
t = 1
t = n/2
t = n
Interactive Tracking Mode
click-prompt
Interactions
box-prompt
boxes
text-prompt
Grounding-DINO
SAM
S^a
Multi-Object Tracking
DeAOT
Add Reference Frame
T^a
Automatic Tracking Mode
S^n
Segment Everything
SAM
Object of Interest Segmentation
text-prompt
Grounding-DINO
boxes
CMR
N^n
T^n
DeAOT
Add Reference Frame

- **SAM-Track (Segment and Track Anything):** Applies SAM to the XMem video segmentation model, achieving an interactive video object segmentation model.
- **Pipeline**
  - **Multimodal Interaction:** Users can select the target through clicking, drawing, or text input.
  - **Automatic Tracking:** SAM-Track, combined with DeAOT, automatically tracks multiple objects in the video.
  - **Enhanced Semantic Understanding:** With Grounding-DINO, SAM-Track supports object selection based on natural language.

Cheng Y, Li L, Xu Y, et al. Segment and track anything[J]. arXiv preprint arXiv:2305.06558, 2023.

ACCV 2024

Machines: LVM-based Trackers


[Human question]:
Who will win the race?
Please track the object.
LoRA
Large Vision-Language Model
[TrackGPT response]:
Sure, I track the object.
Text De-Tokenizer
I_0
I_r
Rethinking
NO
Meet with Purport?
Q_p
$[Q_R, Q_t]$
Q_R
MLP Proj
[PO][TK]
S_p
Visual Encoder
Visual Decoder
S_p
Q_{t+1}
I_t
Video: X_{video}
Tracking Result: m_t

- **TrackGPT:** Proposes a new object tracking task—**Instruction Tracking**. Tracker autonomously reasons and tracks objects in video based on implicit instructions, rather than relying on explicit bounding boxes or language descriptions.
- **Core Design**:
  - **Self-Reasoning:** Utilizes LVLM to understand implicit instructions and reason about the target object.
  - **Cross-Frame Propagation Mechanism:** Adapt to appearance changes.
  - **Rethinking Mechanism:** When the tracking results do not align with the instructions, TrackGPT automatically adjusts and updates the tracking process.

Zhu J, Cheng Z Q, He J Y, et al. Tracking with human-intent reasoning[J]. arXiv preprint arXiv:2312.17448, 2023.

ACCV 2024

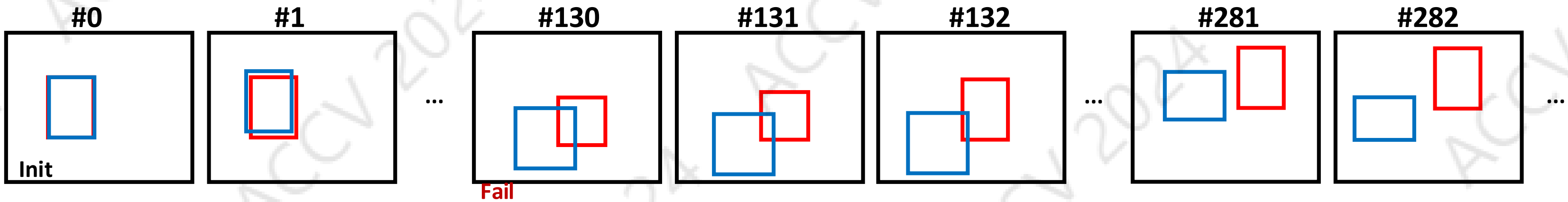
Machine-to-Machine Evaluation

Algorithms are evaluated against benchmarks by comparing their outputs with other algorithms.

ACCV 2024

Machine-to-Machine Evaluation

- **Evaluation Mechanism: OPE**



Traditional OPE Mechanism

➢ **OPE (One-Pass Evaluation):** One-Pass Evaluation (OPE) is a fundamental evaluation method in SOT. It evaluates a tracker by **initializing it in the first frame and letting it run through the entire sequence without reinitialization**.

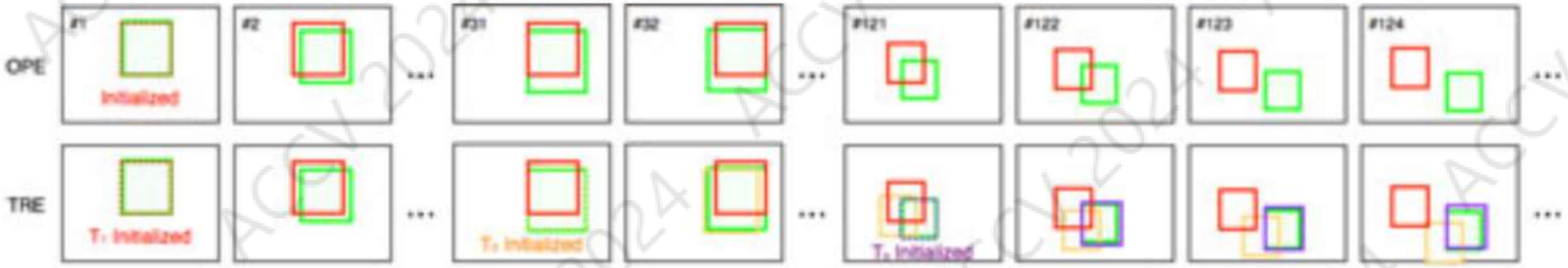
  - **Objective:** Measure the accuracy and robustness of tracking algorithms by allowing the tracker to operate without any manual reinitialization.
  - **Limitations:**
    - ☐ **Influence of initialization**: A poor initialization may cause significant variations in results, as **different starting points could lead to major differences** in tracking performance. → **TRE, SRE**
    - ☐ **Tracking failure**: When tracking fails, the tracker **continues to lose the target** for the remainder of the sequence, providing no meaningful insights after failure. → **Restart Mechanism**

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

ACCV 2024

Machine-to-Machine Evaluation

- **Evaluation Mechanism: TRE**



➢ **Temporal Robustness Evaluation (TRE)** tests the robustness of tracking algorithms by **reinitializing the tracker at different starting points** throughout the sequence. This is done to simulate varying temporal conditions.

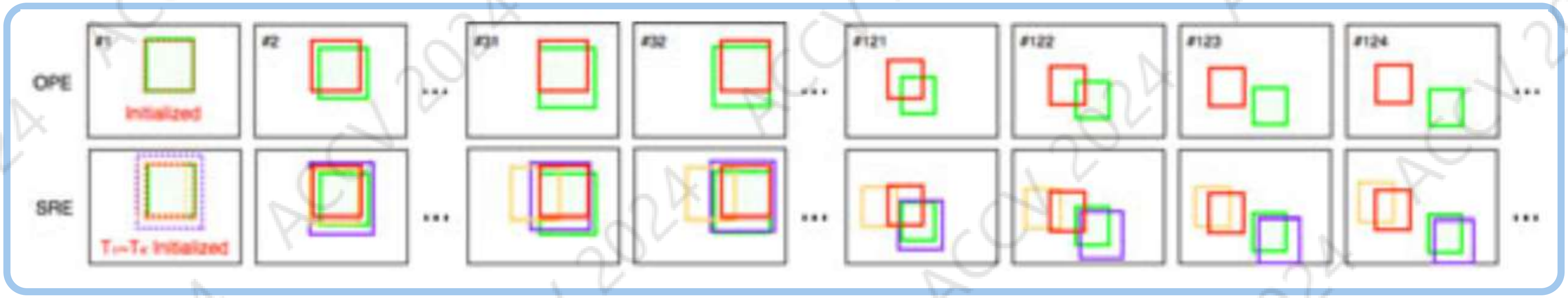
- **Objective:** Evaluate how well a tracker can **handle different temporal conditions** by restarting tracking from multiple frames.
- **TRE Key Metrics:**
  - ☐ **Average Performance:** The tracker is evaluated on the entire sequence by measuring precision and success from different starting points.
  - ☐ **Consistency:** Measures how consistently the tracker performs when initialized at different times within the same sequence.
- **Applications:** TRE is valuable for testing how well tracking algorithms can recover from failures or adapt to changes in target appearance over time.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

ACCV 2024

Machine-to-Machine Evaluation

- ## Evaluation Mechanism: SRE



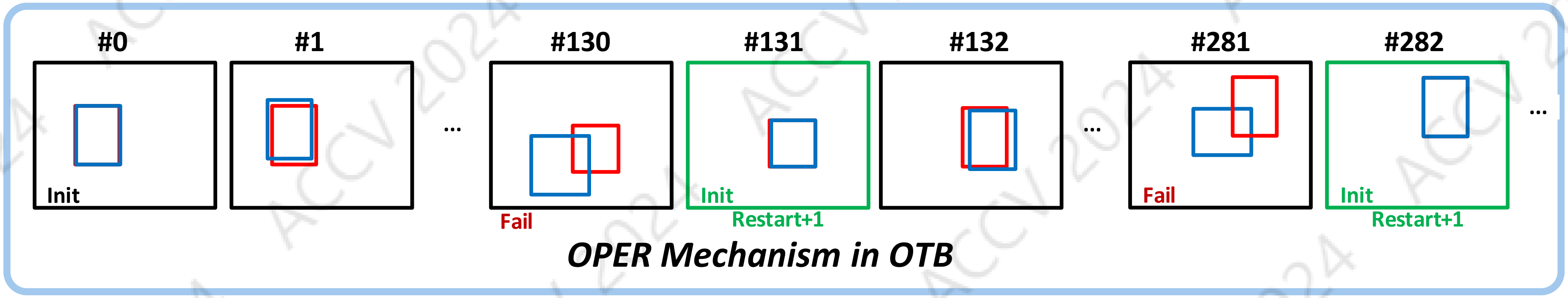
- **Spatial Robustness Evaluation (SRE)** involves **perturbing the initial position or scale of the target** in the first frame. This tests how well the tracker can **handle variations in the initial spatial position or size**.
  - **Objective:** Evaluate a tracker's robustness to **spatial changes** by starting tracking with slight errors in the initial bounding box.
  - **SRE Key Metrics:**
    - ☐ **Tracking Accuracy:** Measures how well the tracker adapts to slight errors in the starting bounding box.
    - ☐ **Resilience to Perturbations:** Evaluates the tracker's ability to handle errors in position or scale during initialization.
  - **Applications:** SRE is particularly useful for testing robustness to inaccuracies in manual annotations or initial target detection errors.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

ACCV 2024

Machine-to-Machine Evaluation

- **Evaluation Mechanism: Restart for OTB Benchmark**



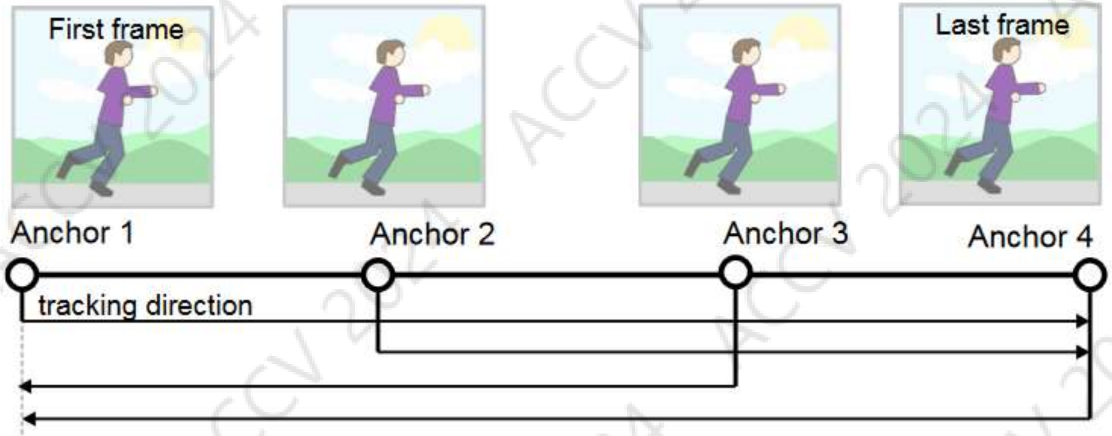
OPER Mechanism in OTB

- **Restart Mechanism** is designed to **reinitialize a tracker when tracking failure is detected**, offering a solution to **prolonged tracking failures**. Originally implemented in OTB, this mechanism has been further developed and adapted to improve evaluation consistency across different tracking scenarios.
  - **OPER (OPE with Restart):** OPER reinitializes the tracking algorithm upon failure and resets the target in the next frame. This ensures that **tracking performance is not unfairly penalized by cumulative errors**, as the evaluation will continue with re-initialized target information.
  - **SRER (SRE with Restart):** SRER similarly handles spatial challenges by reinitializing the tracker in spatially robust environments. It is particularly useful in long-term tracking scenarios where objects reappear after disappearing from the frame.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

Machine-to-Machine Evaluation

- **Evaluation Mechanism: Restart for VOT Challenge**



Anchor-based Evaluation Mechanism for VOT2020

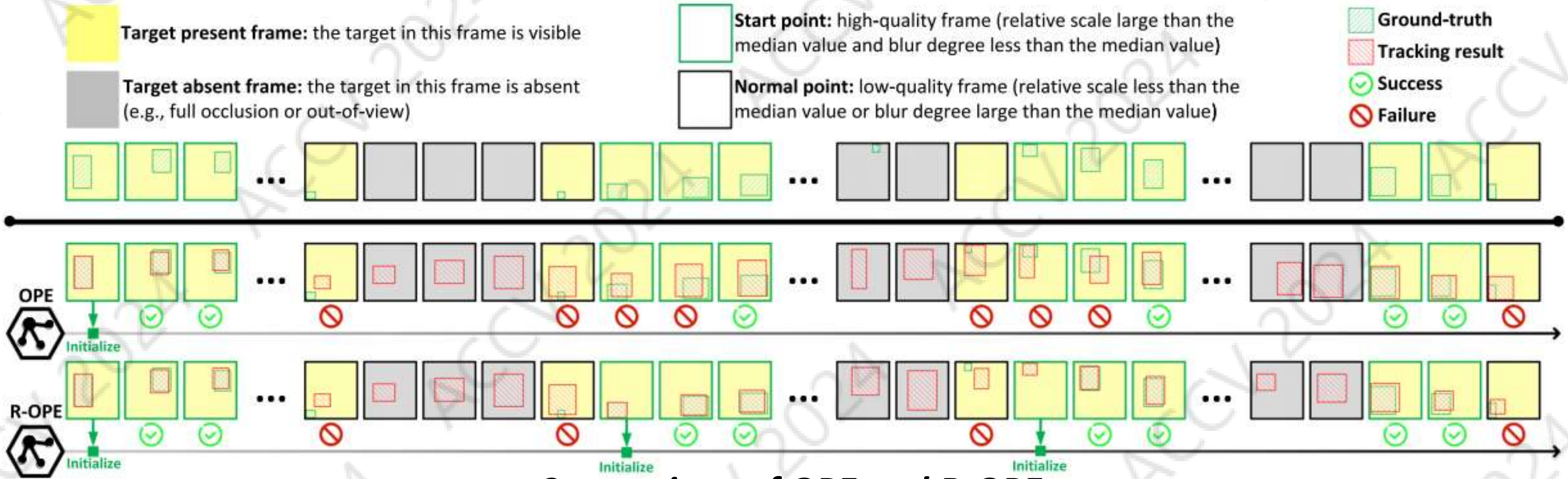
- ➢ Traditional reset mechanism for VOT is similar to OPER, but this mechanism can **introduce causal correlations** between the first reset and subsequent ones, affecting the evaluation fairness.
- ➢ The reset mechanism is replaced by **initialization points** called **anchors**. These are equally spaced along the sequence, removing tracker dependence and ensuring consistency in evaluation.
  - **Anchor Placement:** Anchors are placed ***every Δanc frames*** throughout the sequence. The first and last anchors are at the start and end of the sequence.
  - **Tracking Direction:** A tracker is run from each anchor either forward or backward to ensure the longest possible sub-sequence is used for evaluation.

https://www.votchallenge.net/

ACCV 2024

Machine-to-Machine Evaluation

- ## Evaluation Mechanism: Restart for VideoCube


Target present frame: the target in this frame is visible
Target absent frame: the target in this frame is absent (e.g., full occlusion or out-of-view)
Start point: high-quality frame (relative scale large than the median value and blur degree less than the median value)
Normal point: low-quality frame (relative scale less than the median value or blur degree large than the median value)
Ground-truth
Tracking result
Success
Failure
OPE
Initialize
R-OPE
Initialize
Initialize
Initialize

Comparison of OPE and R-OPE

- **R-OPE (Restart-Based OPE)** is a new reset mechanism introduced in 2023 for the VideoCube benchmark, particularly designed to address real-world tracking tasks that involve complex scenarios.
- **Key Concept:**
  - In this mechanism, the tracker is reset not immediately after failure, but at the nearest anchor point (**frame with clear appearance information**).
  - By choosing optimal restart points, R-OPE **avoids repeatedly initializing in problematic regions of the video**.

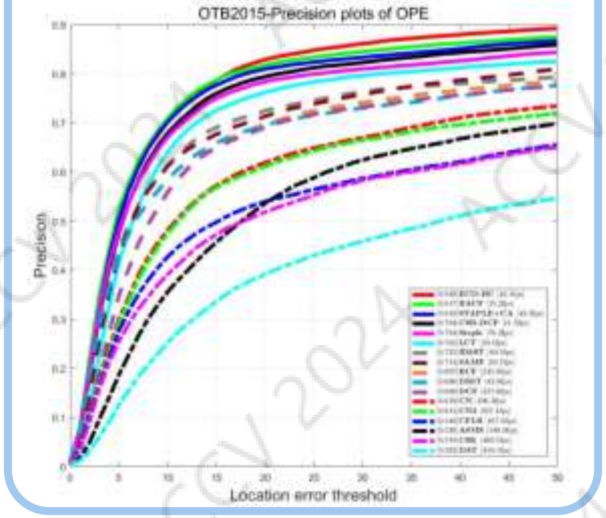
***S. Hu, X. Zhao#**, L. Huang, et al., “Global instance tracking: Locating target more like humans,” IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

Machine-to-Machine Evaluation

- ## Evaluation Metris: Precision

$$P(G) = \frac{1}{|G|}\sum_{s_i \in G}\frac{1}{|s_i|}\left|\{F_i : d_c \le \theta_d\}\right|$$

$$d_c = \left\| c_p - c_g \right\|_2$$



- **Precision (PRE):** Precision is one of the most commonly used evaluation metrics in single object tracking tasks, primarily used to measure the accuracy of the predicted result. It reflects **how closely the predicted target position matches the actual target position in each frame.**
- **Calculation Method:** Precision is typically calculated by measuring the **Euclidean distance between the predicted target center and the ground truth center**. If this distance is smaller than a predefined threshold, the frame is considered as correctly tracked. The proportion of such frames over the total number of frames gives the precision.
  - **Common Threshold:** A typical threshold for tracking tasks is **20 pixels**, meaning if the distance between the predicted center and the true center is less than 20 pixels, the tracking is deemed successful for that frame.

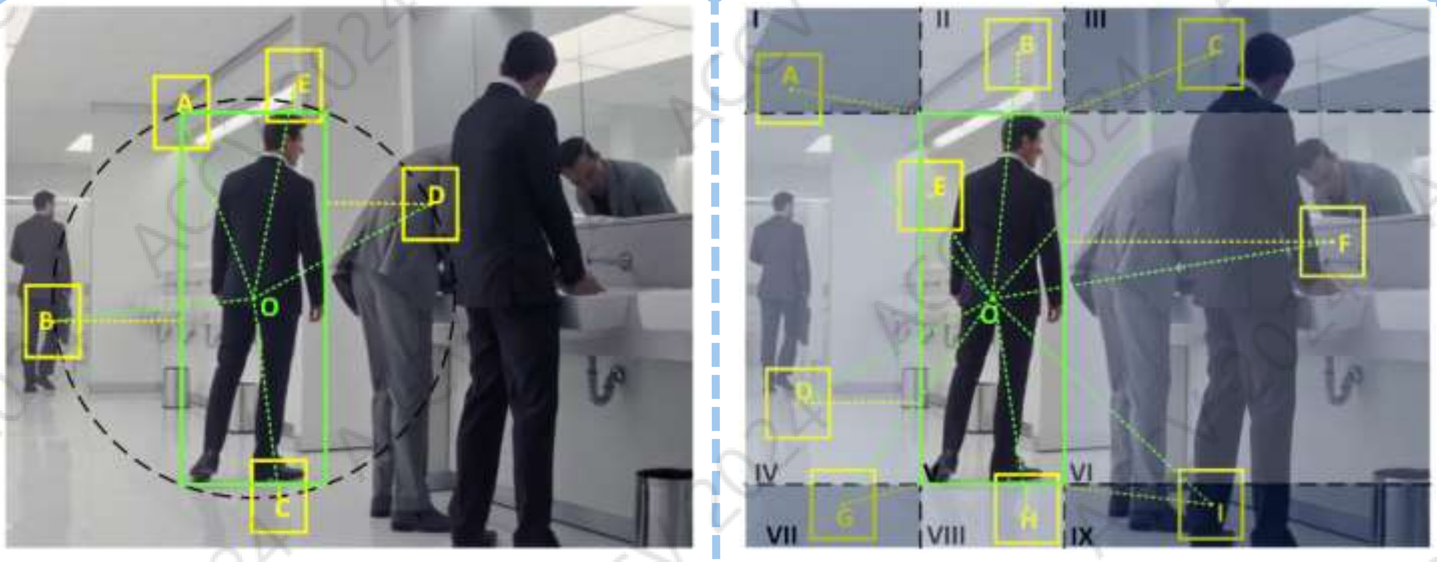
Limitations:
In cases where the target's size changes significantly or has irregular shapes, evaluating solely by center point distance can introduce biases.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

ACCV 2024

Machine-to-Machine Evaluation

- **Evaluation Metris: Precision**



PRE: A=B=C=D=E

$$P(G) = \frac{1}{|G|}\sum_{s_i \in G}\frac{1}{|s_i|}\left|\{F_i : d_c \leq \theta_d\}\right|$$

$$d_c = \left\| c_p - c_g \right\|_2$$

N-PRE:E is the best

$$P'(G) = \frac{1}{|G|}\sum_{s_i \in G}\frac{1}{|s_i|}\left|\{F_i : N'(d_c) \leq \theta'_d\}\right|$$

$$N'(d_c) = \frac{d'_c}{\max(\{d'_i \mid i \in F_i\})}$$

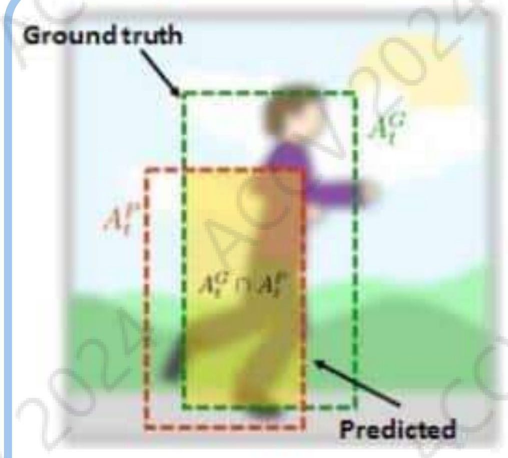
➢ **Normalized Precision (N-PRE):** Normalized precision is calculated by **normalizing the center error** as a ratio of the target's scale. Specifically, **the target width and frame resolution** are combined as the normalization factor, and the Euclidean distance between the predicted center and the groundtruth center is normalized. A normalized threshold is then used to determine if the tracking is accurate.

*S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

ACCV 2024

Machine-to-Machine Evaluation

- **Evaluation Metris: Success Rate**



$$S(G) = \frac{1}{|G|}\sum_{s_i \in G}\frac{1}{|s_i|}|\{F_i : s_i \geq \theta_i\}|$$

$$s_i = \Omega(p_i, g_i) = \frac{p_i \cap g_i}{p_i \cup g_i}$$

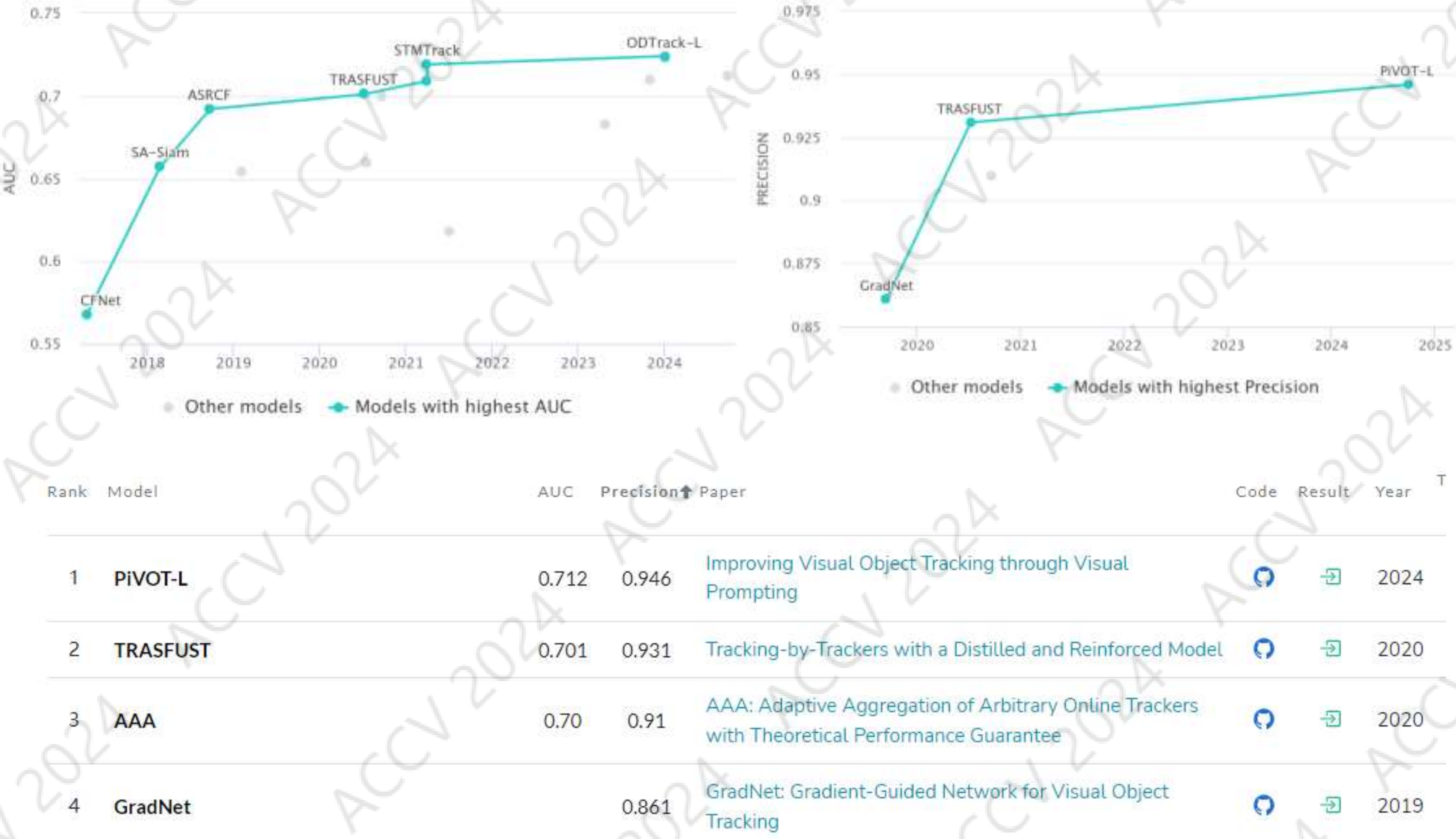
Limitations:
In cases where the predicted and ground truth bounding boxes have little to no overlap, IoU cannot capture the spatial relationship between them, leading to a zero score.

- **Success Rate (SR):** SR is one of the key evaluation metrics used in single object tracking tasks to assess the overall performance of tracking algorithms. Unlike PRE, which focuses on the accuracy of the center point, success rate evaluates **how well the predicted bounding box overlaps with the ground truth bounding box**, providing a more comprehensive measure of the tracking performance in terms of object detection and position tracking.
- **Calculation Method:** Success rate is determined by calculating the **Intersection over Union (IoU)** between the predicted bounding box and the ground truth bounding box. IoU measures the overlap between the two bounding boxes as a ratio of their intersection area to their union area. If the IoU is greater than a predefined threshold (**typically 0.5**), the frame is considered to be successfully tracked.

Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2411-2418.

Machine-to-Machine Evaluation

- **State-of-the-art Results: OTB100**

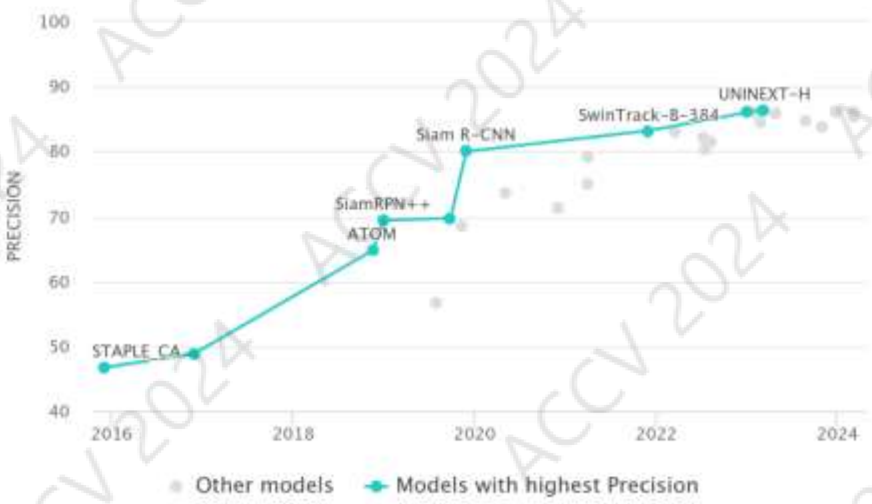

0.75
0.7
0.65
0.6
0.55
AUC
CFNet
SA-Siam
ASRCF
TRASFUST
STMTrack
ODTrack-L
2018
2019
2020
2021
2022
2023
2024
Other models
Models with highest AUC
0.975
0.95
0.925
0.9
0.875
0.85
PRECISION
GradNet
TRASFUST
PiVOT-L
2020
2021
2022
2023
2024
2025
Other models
Models with highest Precision

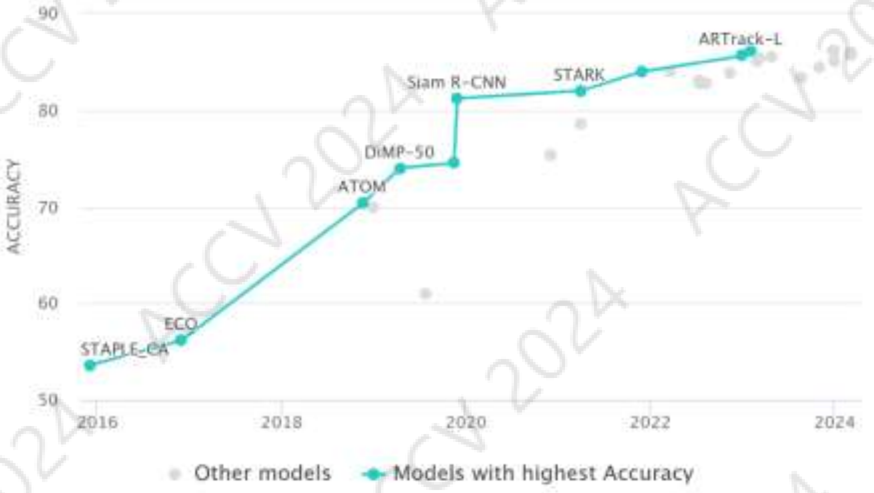
Rank	Model	AUC	Precision↑	Paper	Code	Result	Year	T
1	**PiVOT-L**	0.712	0.946	Improving Visual Object Tracking through Visual Prompting			2024	
2	**TRASFUST**	0.701	0.931	Tracking-by-Trackers with a Distilled and Reinforced Model			2020	
3	**AAA**	0.70	0.91	AAA: Adaptive Aggregation of Arbitrary Online Trackers with Theoretical Performance Guarantee			2020	
4	**GradNet**		0.861	GradNet: Gradient-Guided Network for Visual Object Tracking			2019	

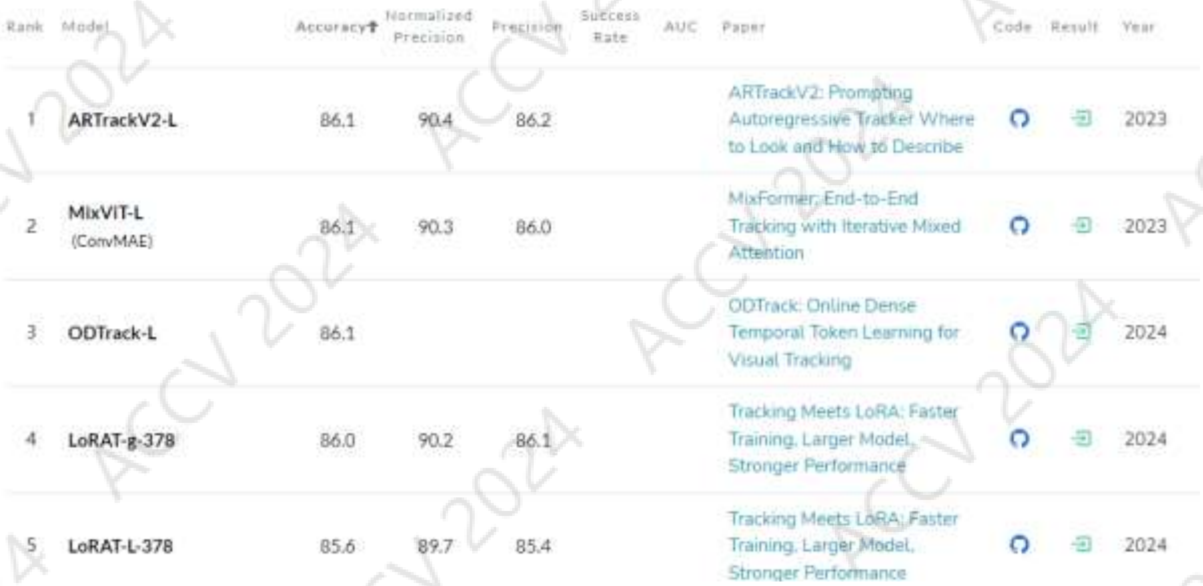
https://paperswithcode.com/sota/visual-object-tracking-on-otb-2015

Machine-to-Machine Evaluation

- **State-of-the-art Results: TrackingNet**







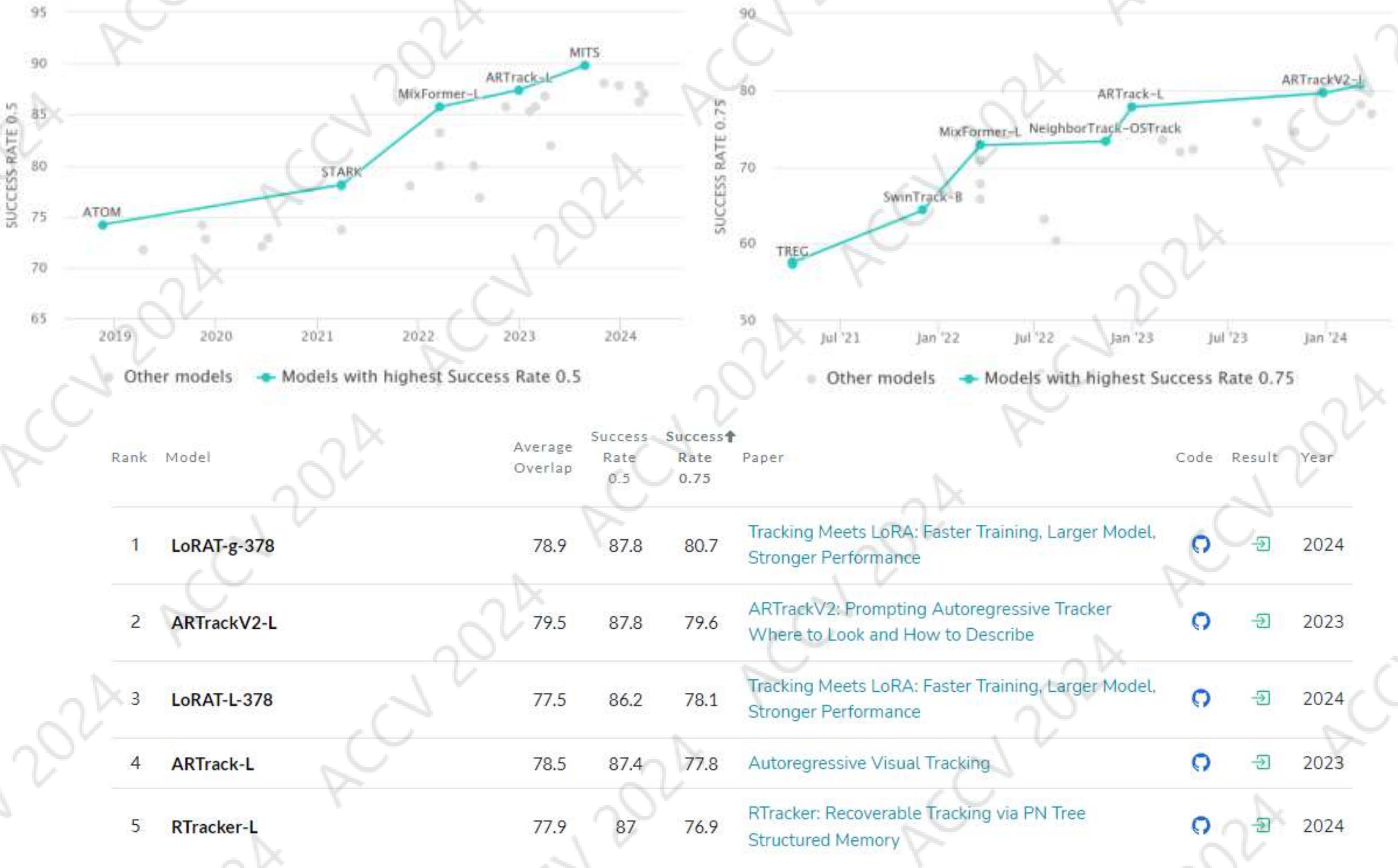
Rank	Model	Accuracy	Normalized Precision	Precision	Success Rate	AUC	Paper	Code	Result	Year
1	ARTrackV2-L	86.1	90.4	86.2			ARTrackV2: Prompting Autoregressive Tracker Where to Look and How to Describe			2023
2	MixViT-L (ConvMAE)	86.1	90.3	86.0			MixFormer: End-to-End Tracking with Iterative Mixed Attention			2023
3	ODTrack-L	86.1					ODTrack: Online Dense Temporal Token Learning for Visual Tracking			2024
4	LoRAT-g-378	86.0	90.2	86.1			Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance			2024
5	LoRAT-L-378	85.6	89.7	85.4			Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance			2024

https://paperswithcode.com/sota/visual-object-tracking-on-trackingnet

ACCV 2024

Machine-to-Machine Evaluation

- **State-of-the-art Results: GOT-10k**


SUCCESS RATE 0.5
95
90
85
80
75
70
65
2019
2020
2021
2022
2023
2024
ATOM
STARK
MixFormer-L
ARTrack-L
MITS
Other models
Models with highest Success Rate 0.5
SUCCESS RATE 0.75
90
80
70
60
50
Jul '21
Jan '22
Jul '22
Jan '23
Jul '23
Jan '24
TREG
SwinTrack-B
MixFormer-L
NeighborTrack-OSTrack
ARTrack-L
ARTrackV2-L
Other models
Models with highest Success Rate 0.75

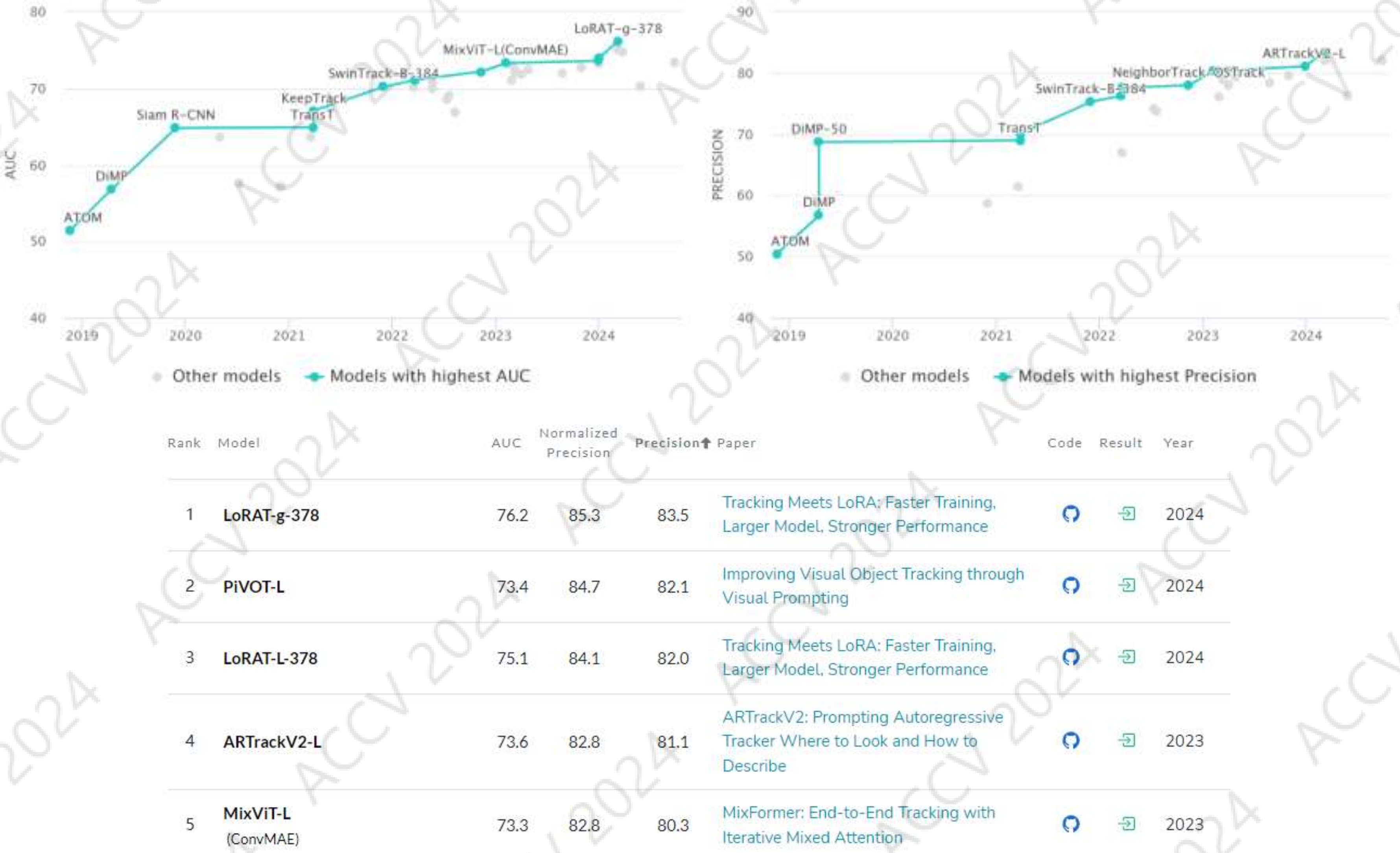
Rank	Model	Average Overlap	Success Rate 0.5	Success Rate 0.75 ↑	Paper	Code	Result	Year
1	**LoRAT-g-378**	78.9	87.8	80.7	Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance	●	→	2024
2	**ARTrackV2-L**	79.5	87.8	79.6	ARTrackV2: Prompting Autoregressive Tracker Where to Look and How to Describe	●	→	2023
3	**LoRAT-L-378**	77.5	86.2	78.1	Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance	●	→	2024
4	**ARTrack-L**	78.5	87.4	77.8	Autoregressive Visual Tracking	●	→	2023
5	**RTracker-L**	77.9	87	76.9	RTracker: Recoverable Tracking via PN Tree Structured Memory	●	→	2024

https://paperswithcode.com/sota/visual-object-tracking-on-got-10k

ACCV 2024

Machine-to-Machine Evaluation

- **State-of-the-art Results: LaSOT**


AUC
80
70
60
50
40
2019
2020
2021
2022
2023
2024
ATOM
DiMP
Siam R-CNN
KeepTrack
TransT
SwinTrack-B-384
MixViT-L(ConvMAE)
LoRAT-g-378
Other models
Models with highest AUC
PRECISION
90
80
70
60
50
40
2019
2020
2021
2022
2023
2024
ATOM
DiMP
DiMP-50
TransT
SwinTrack-B-384
NeighborTrack
OSTrack
ARTrackV2-L
Other models
Models with highest Precision

Rank	Model	AUC	Normalized Precision	Precision↑	Paper	Code	Result	Year
1	**LoRAT-g-378**	76.2	85.3	83.5	Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance			2024
2	**PiVOT-L**	73.4	84.7	82.1	Improving Visual Object Tracking through Visual Prompting			2024
3	**LoRAT-L-378**	75.1	84.1	82.0	Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance			2024
4	**ARTrackV2-L**	73.6	82.8	81.1	ARTrackV2: Prompting Autoregressive Tracker Where to Look and How to Describe			2023
5	**MixViT-L** (ConvMAE)	73.3	82.8	80.3	MixFormer: End-to-End Tracking with Iterative Mixed Attention			2023

https://paperswithcode.com/sota/visual-object-tracking-on-lasot

ACCV 2024

Conclusion

- **Machines:**
  - Traditional Trackers
  - Correlation Filter Based Trackers
  - Siamese Neural Network Based Trackers
  - One-stream Trackers
  - LVM-based Trackers
  - **Impact:** These advancements highlight the shift from basic motion modeling to sophisticated, context-aware algorithms **that bring machine performance closer to human-like tracking abilities**.

ACCV 2024

Conclusion

Machine-to-Machine Evaluation:

- Machine-to-machine evaluation has **long served as the primary method** for assessing tracking performance, relying on benchmarks that compare algorithmic outputs with ground-truth data in controlled environments.
- This approach focuses on metrics like accuracy, robustness, and computational efficiency, which are **valuable for assessing fundamental tracking capabilities**.

ACCV 2024

Time Break

About 20 Minutes

ACCV 2024

CONTENTS

- Introduction
- Part 1. Visual Object Tracking Task
- Part 2. Experimental Environments
- Part 3. Algorithms and Traditional Machine-Machine Comparisons
- **Part 4. Human Visual Abilities and Visual Turing Test**
  - **Humans**
  - **Visual Turing Test**
- Trends and Future Directions

ACCV 2024

Scan to download this tutorial PPT

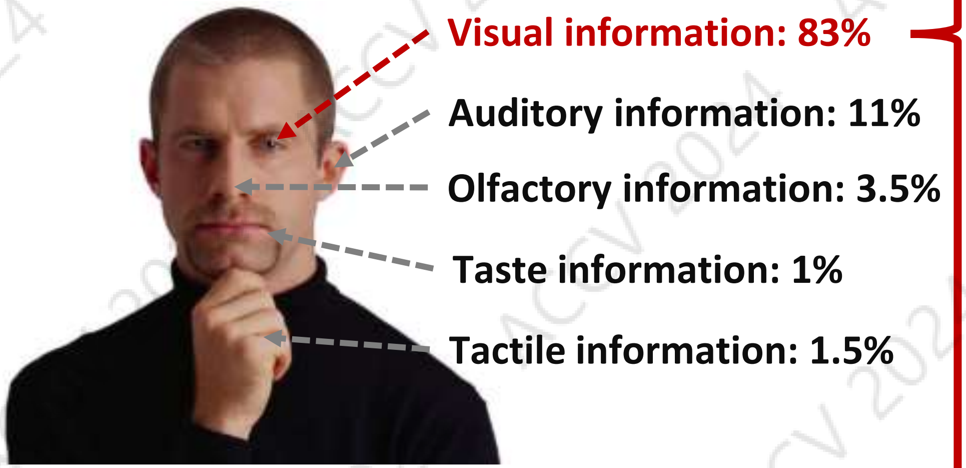
Humans

We will learn some basic information about human visual theories and abilities.

ACCV 2024

Humans: Visual Theories

- **Introduction to Visual Theories**



Visual information: 83%

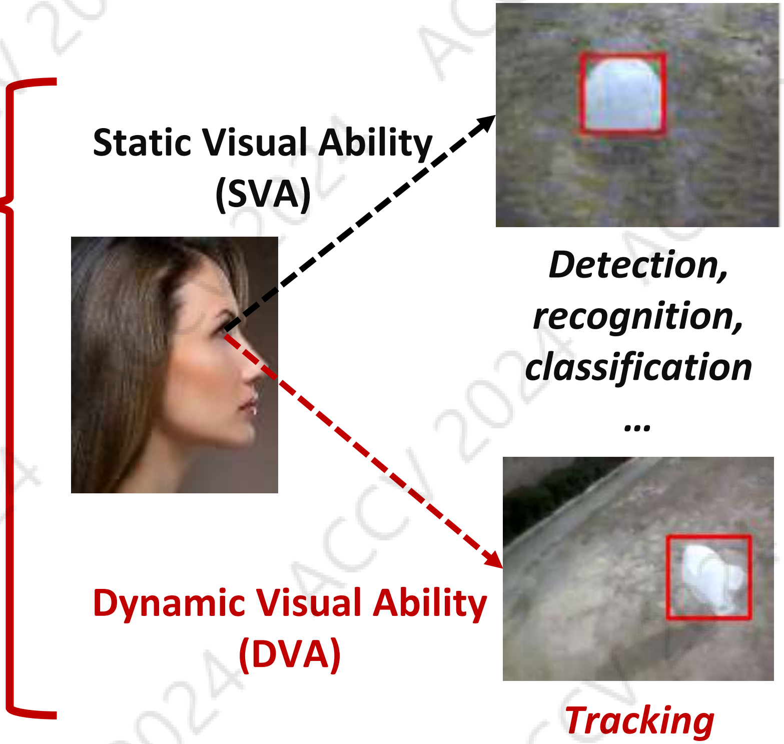
Auditory information: 11%

Olfactory information: 3.5%

Taste information: 1%

Tactile information: 1.5%

Humans are "visual animals"



Static Visual Ability (SVA)

Detection, recognition, classification …

Dynamic Visual Ability (DVA)

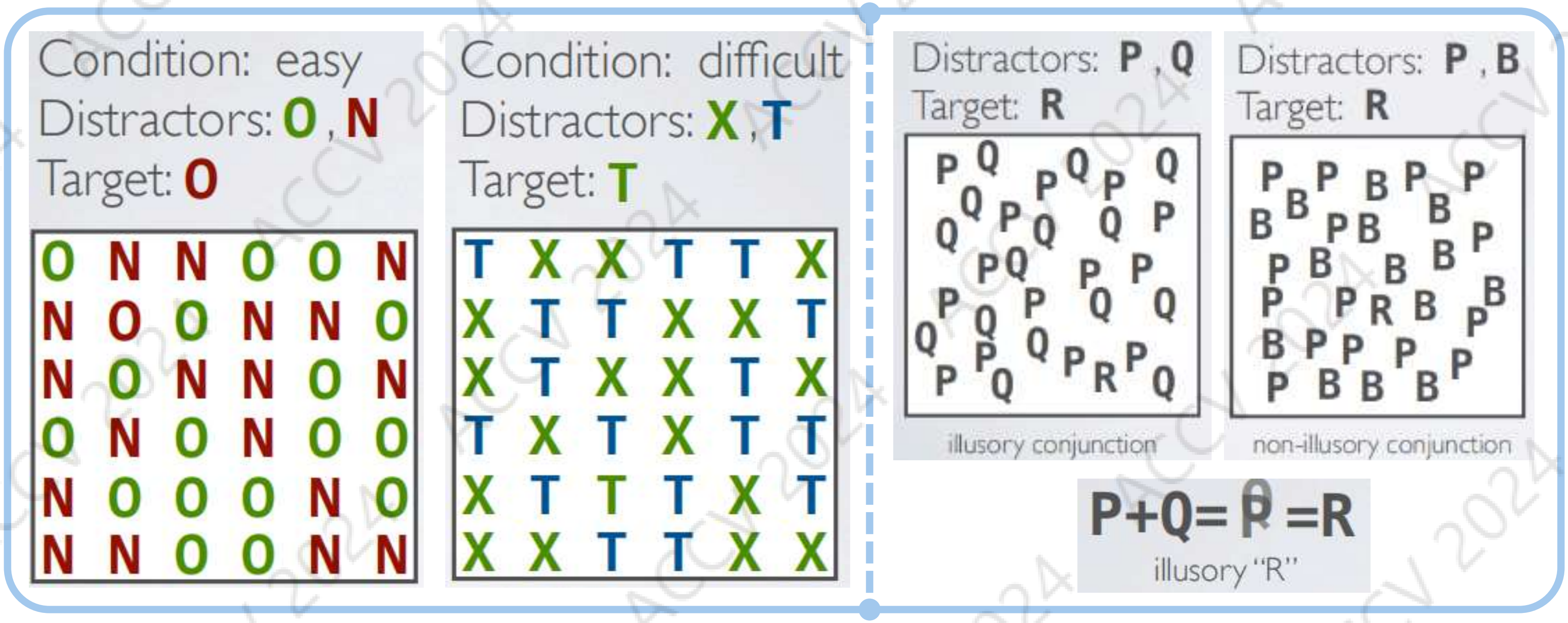
Tracking

- Visual theories offer insights into **how humans process and interpret visual information**, providing a foundation for improving algorithms.
- Theories such as **Feature Integration Theory, Recognition-by-Components**, and **Visual Computation** have inspired the development of advanced computer vision systems.

ACCV 2024

Humans: Visual Theories

- **Feature Integration Theory (1980)**


Condition: easy
Distractors: O , N
Target: O
O N N O O N
N O O N N O
N O N N O N
O N O N O O
N O O O N O
N N O O N N
Condition: difficult
Distractors: X , T
Target: T
T X X T T X
X T T X X T
X T X X T X
T X T X T T
X T T T X T
X X T T X X
Distractors: P , Q
Target: R
illusory conjunction
Distractors: P , B
Target: R
non-illusory conjunction
P+Q= R =R
illusory "R"

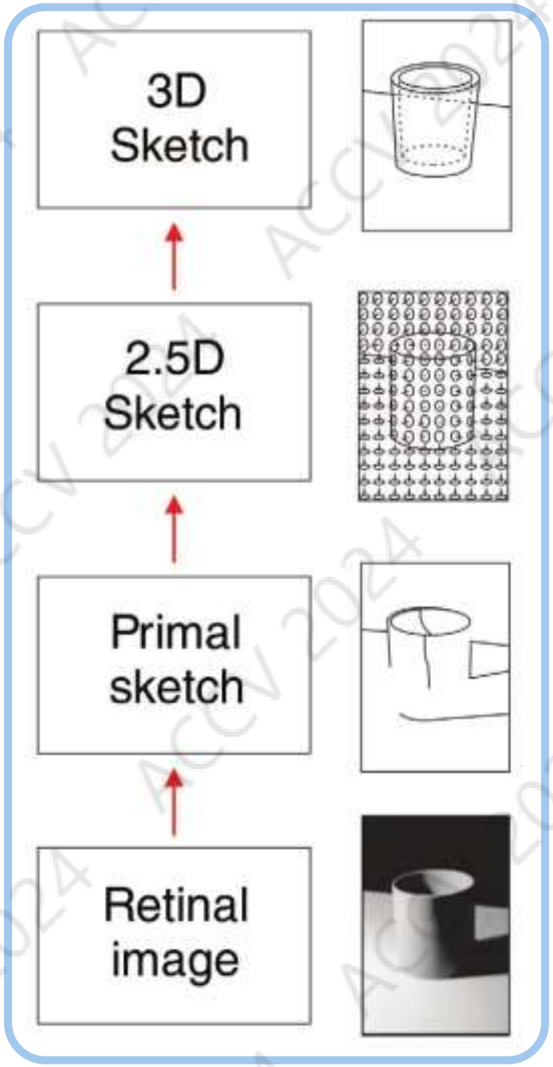
- This theory explains how humans **combine separate visual features** (such as color, shape, size) **to form a coherent object perception**. It suggests that the brain **processes simple features in parallel** and integrates them into a unified perception during focused attention.
- **Applications:** Feature-based tracking algorithms mimic this theory by **extracting multiple object characteristics** (such as color, texture, shape) to maintain robust tracking performance, especially in complex environments.

TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive psychology, 1980, 12(1): 97-136.

ACCV 2024

Humans: Visual Theories

- **Visual Computation Theory (1982)**



- Describes how the brain processes visual information through a series of **hierarchical stages**, starting from a basic edge detection to the construction of complex, 3D object representations. The theory breaks down vision into three main levels:
  - **Raw primal sketch**: Initial edge and texture information.
  - **2.5D sketch**: Intermediate-level representation of objects' positions and orientation.
  - **3D representation**: Full object understanding for recognition and interaction.
- **Applications:** This theory underlies many modern tracking algorithms, where visual data is processed hierarchically, starting from **low-level features** (e.g., edges, textures) to **high-level representations** (e.g., object shapes, motions).

Marr D. Vision: A computational investigation into the human representation and processing of visual information[M]. MIT press, 2010.

ACCV 2024

Humans: Visual Theories

- **Recognition-by-Components Theory (1987)**

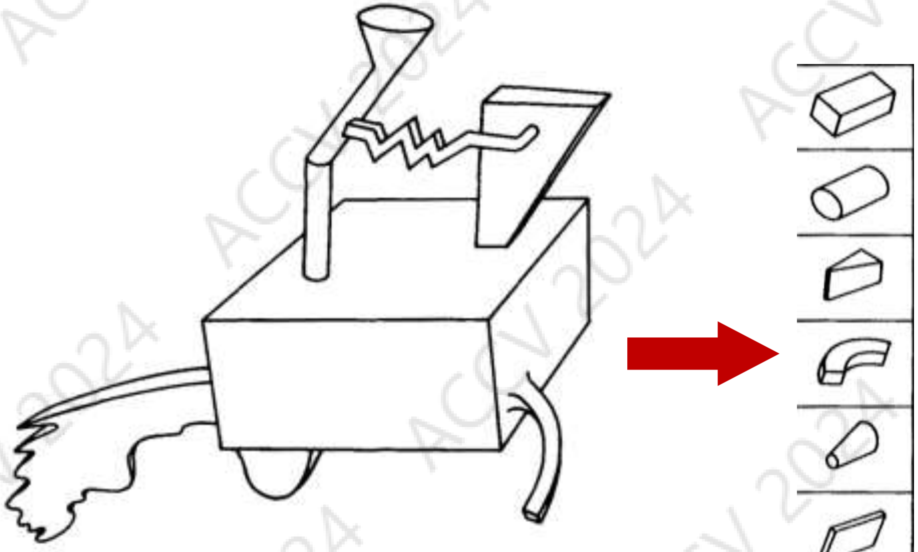
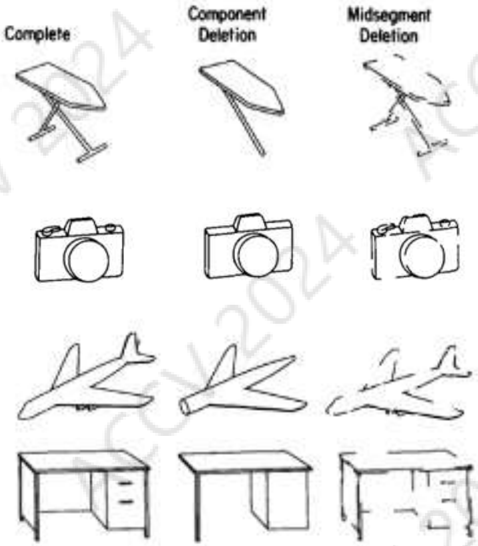


Figure 1. A do-it-yourself object. (There is strong consensus in the segmentation loci of this configuration and in the description of its parts.)



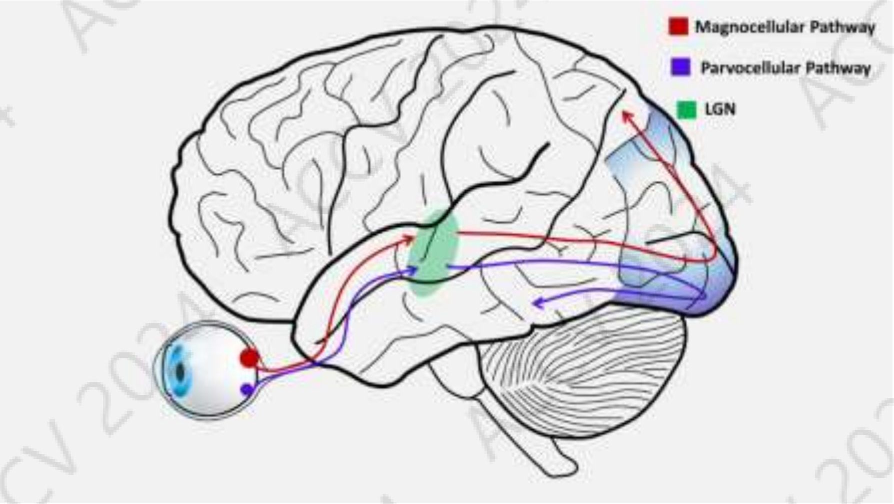
- ➢ Suggests that humans recognize objects by breaking them down into basic geometric shapes, called "**geons**." These geons form the building blocks of object recognition. The theory argues that recognizing the components is sufficient for object identification, even if some parts of the object are obscured.
- ➢ **Applications:** Inspired object segmentation techniques in tracking, where objects are broken down into parts for **more accurate identification and tracking**.

Biederman I. Recognition-by-components: a theory of human image understanding[J]. Psychological review, 1987, 94(2): 115.

ACCV 2024

Humans: Visual Ability

- **Overview of Human Visual Capabilities**



- ➢ The **magnocellular** (M cell) pathway carries information about **large, fast things** (low spatial frequency; high temporal frequency) and is **colorblind**.
- ➢ The **parvocellular** (P Cell) pathway carries information about **small, slow, colorful things** (high spatial frequency information; low temporal frequency information).

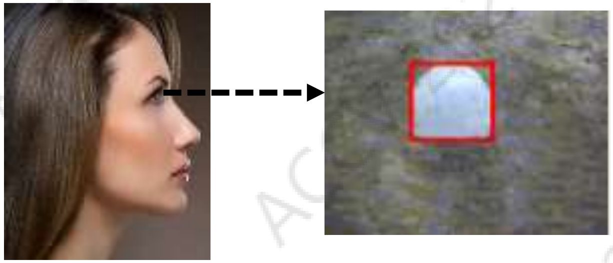
- ➢ Pioneering research from a neurophysiological has allowed distinction between the two main types of visual acuity:
  - **Static Visual Ability:** The ability to perceive and interpret **stationary or slow-moving objects**, whose basic neural support is the **parvocellular system.**
  - **Dynamic Visual Ability:** The ability to perceive and track **fast-moving objects** or predict their trajectories, whose basic neural support is the **magnocellular system.**

JW M, Ludvigh E. The effect of relative motion on visual acuity[J]. Survey of Ophthalmology, 1962, 7: 83-116.

ACCV 2024

Humans: Visual Ability

- **Static Visual Capability**



Detection, recognition, classification …

➢ **Definition:** SVA is defined as the ability to distinguish the details of **static objects** whose image is formed on the retina when the evaluated subject is also **stationary.**

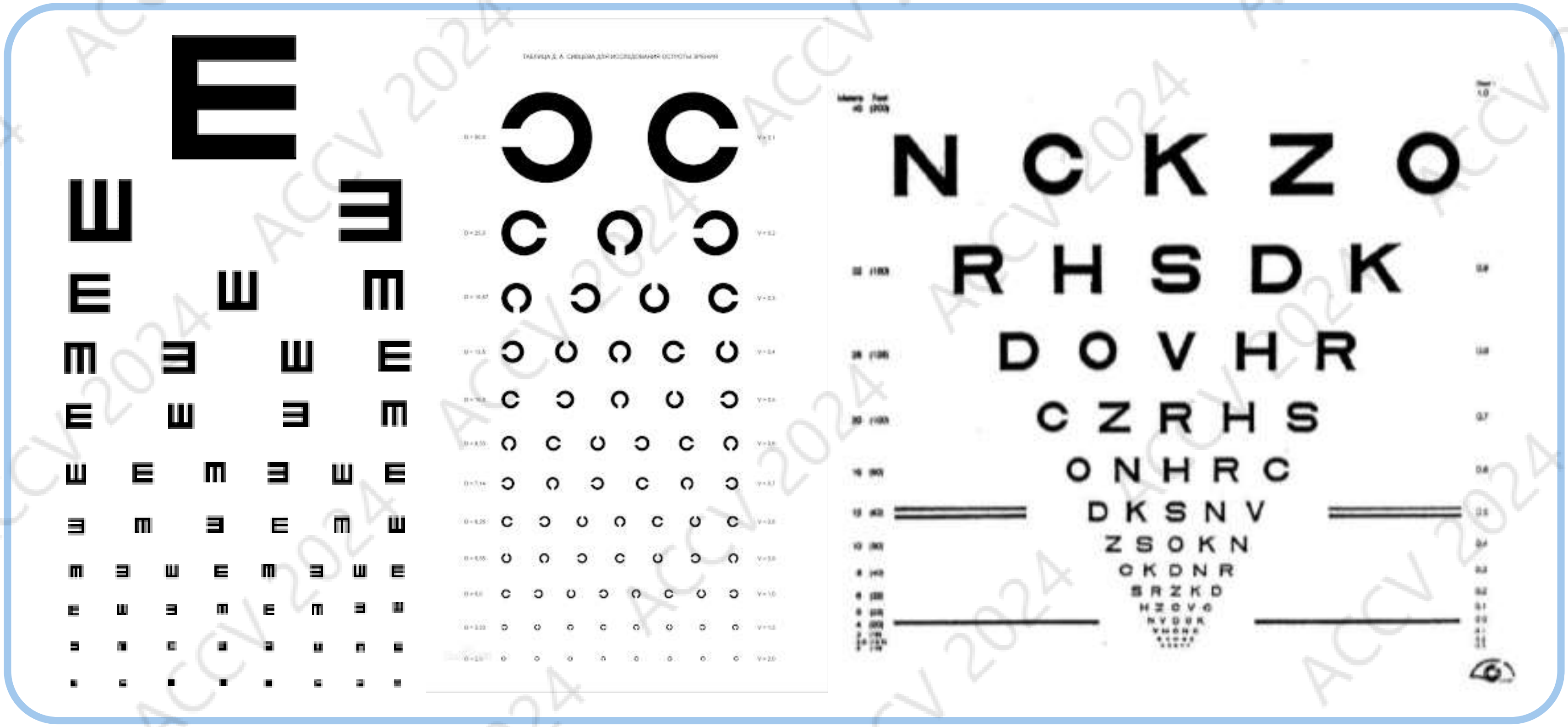
➢ **Key points for measuring SVA:** In assessing this visual ability, some basic thresholds can be considered:

- **Minimum detectable threshold:** ability to perceive the smallest object in the visual field.
- **Minimum resolution threshold**: ability to perceive as separate two objects that are very close together.
- **Minimum perceptible alignment threshold**: refers to the ability to detect the alignment between two discontinuous segments whose ends are very close together.
- **Minimum recognition threshold**: ability to properly identify the shape or orientation of an object (e.g. a letter). This threshold is commonly referred to as visual acuity.

ACCV 2024

Humans: Visual Ability

- **Static Visual Capability**


N C K Z O
R H S D K
D O V H R
C Z R H S
O N H R C
D K S N V
Z S O K N
O K D N R

➢ **Measurement Technology:** The limit of **spatial resolution (the smallest size)** that the subject can visually resolve is reached when he is unable to identify the letters of a row (or perceive the distance between two points or lines or the opening of a ring).

ACCV 2024

Humans: Visual Ability

- ## Static Visual Capability

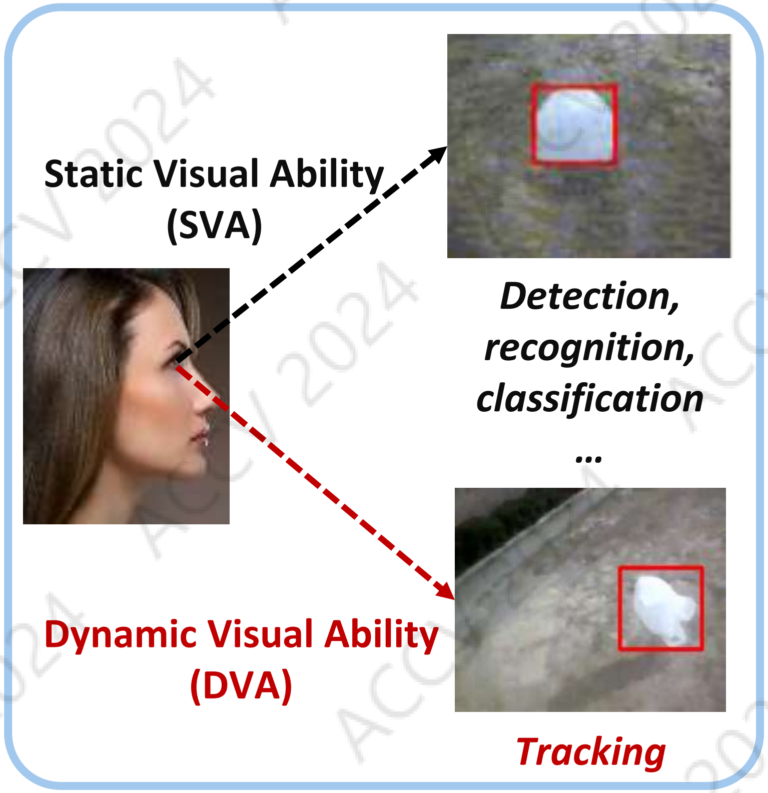
- **Influencing Factors:** Among the subject factors that may **influence** SVA:
  - The most determinants is the **refractive error**, which, in most cases, would require the appropriate optical prescription to achieve normal visual acuity.
  - Another very important element is the **age of the subject**, which is known to lead to anatomical and physiological changes that adversely affect visual perception.
- **Limitations:** There are two limitations that show the inadequacy of measuring only SVA to assess the functioning of the visual system:
  - Many visual stimuli to which we must respond to in real life **are often in motion.**
  - The SVA tests refer to letters or symbols often displayed under conditions of **maximum contrast (black on white),** even though such high level of contrast is seldom observed in the different situations of daily life.

Long, G. M., & Zavod, M. J. (2002). Contrast sensitivity in a dynamic environment: Effects of target conditions and visual impairment.

Humans: Visual Ability

- ## Dynamic Visual Capability

➢ **Definition:** Dynamic visual acuity (DVA) describes the ability to visually **resolve subtle spatial details** of an object when the object, the observer, or both, are **moving**.



Static Visual Ability (SVA)

Detection, recognition, classification …

Dynamic Visual Ability (DVA)

Tracking

➢ **Correlation between DVA and SVA**

- **Static vs. Dynamic:** While dynamic visual capability often builds on static capability, research shows that **having excellent static visual capability does not always mean strong dynamic capability**. Individuals with high static visual acuity might struggle with tracking moving objects.
- **Complementary Abilities: Both capabilities are essential in designing tracking systems** that can handle a wide range of visual tasks, from identifying stationary objects to tracking fast-moving targets in real time.

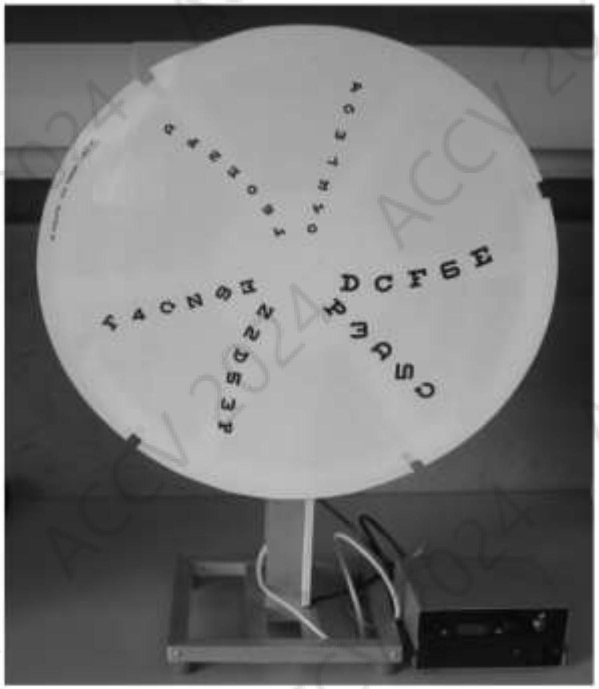
Aznar-Casanova, J. A., Quevedo, L. 1., & Sinnet, S. (2005). The effects of drift and displacement motion on dynamic visual acuity

ACCV 2024

Humans: Visual Ability

- **Dynamic Visual Capability**

➢ **Measurement Technology:** Unfortunately, despite the importance of DVA, specific instruments with proven reliability and validity that enable further research of such ability are **inadequate and chaotic.**



➢ **Bernell's Rotator (1990s):** The dynamic visual acuity values are recorded as a combination of **visual acuity** and **speed** in rpm.

ACCV 2024

Humans: Visual Ability

• Dynamic Visual Capability

- **Measurement Technology:** Unfortunately, despite the importance of DVA, specific instruments with proven reliability and validity that enable further research of such ability are **inadequate and chaotic.**

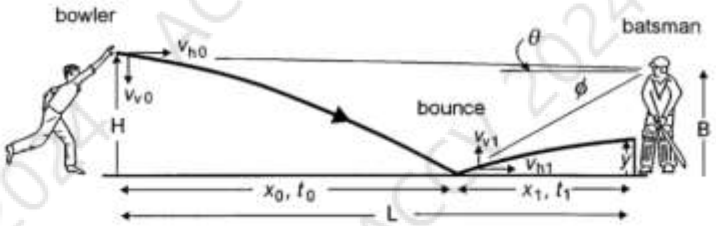
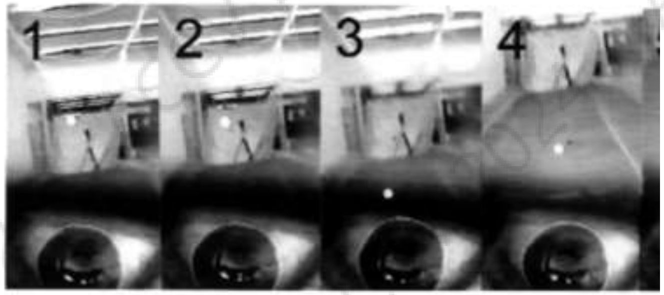
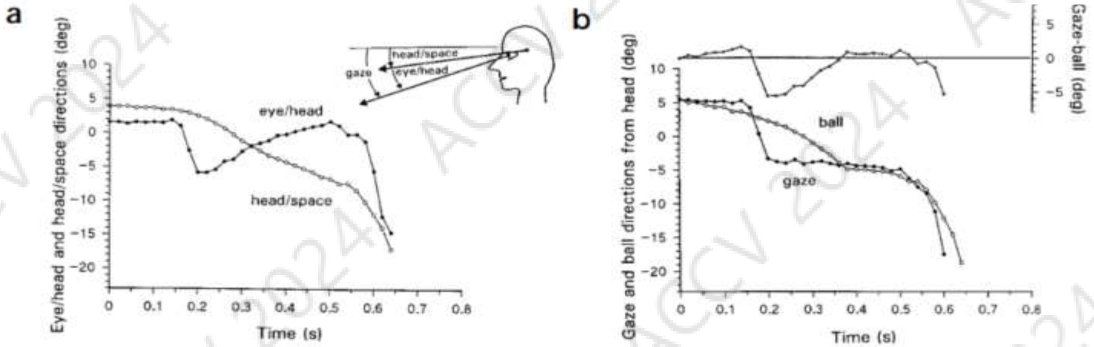


Fig. 1. The physical parameters that determine where and when the ball will reach the batsman, and the measurements he can make to determine the time and point of contact. The batsman's task is to estimate when it will reach him (t_1), and at what height (y). Other definitions are in text.





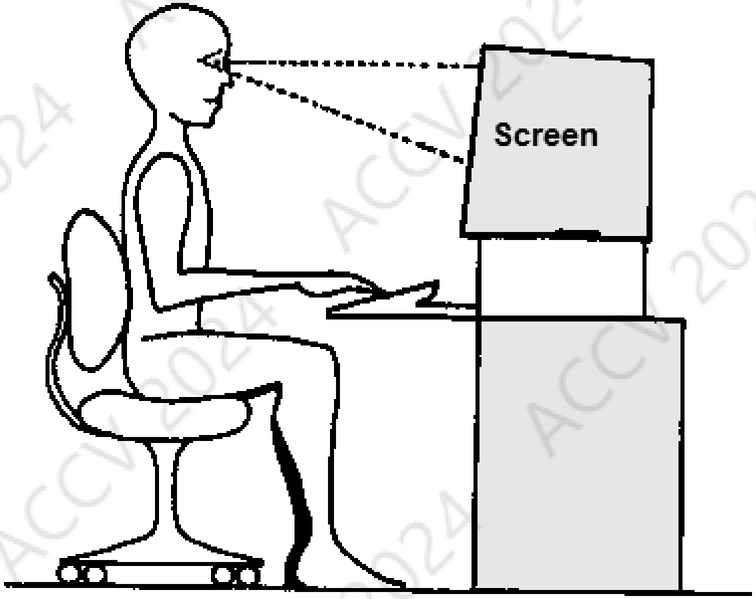
- **Ball game (2000s):** Research in the 2000s focused on **testing athletes' dynamic visual abilities**, primarily in **baseball**, where the athletes' eye movements were recorded and analyzed by cameras.

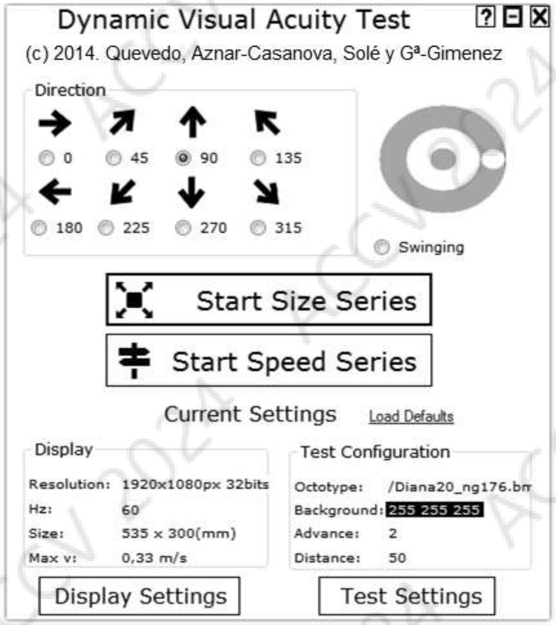
Land M F, McLeod P. From eye movements to actions: how batsmen hit the ball[J]. Nature neuroscience, 2000, 3(12): 1340-1345.
McLeod P, Reed N, Dienes Z. How fielders arrive in time to catch the ball[J]. Nature, 2003, 426(6964): 244-245.

Humans: Visual Ability

- **Dynamic Visual Capability**

➢ **Measurement Technology:** Unfortunately, despite the importance of DVA, specific instruments with proven reliability and validity that enable further research of such ability are **inadequate and chaotic.**




Dynamic Visual Acuity Test

(c) 2014. Quevedo, Aznar-Casanova, Solé y Gª-Gimenez

Direction

○ 0 ○ 45 ● 90 ○ 135

○ 180 ○ 225 ○ 270 ○ 315

○ Swinging

Current Settings

Display

Resolution: 1920x1080px 32bits
Hz: 60
Size: 535 x 300(mm)
Max v: 0,33 m/s

Test Configuration

Octotype: /Diana20_ng176.bm
Background: 255 255 255
Advance: 2
Distance: 50

➢ **DynVA (2010s):** The DynVA is a computer software designed to assess DVA. The researcher can select the optotype to be presented in the two forms of the test: **(a)Size Series; (b) Speed Series.**

Quevedo L, Aznar-Casanova J A, Merindano-Encina D, et al. A novel computer software for the evaluation of dynamic visual acuit

ACCV 2024

Humans: Visual Ability

- ## Dynamic Visual Capability

➢ **Influencing Factors:** Among the subject factors that may **influence** DVA:

- **Age:**
  - ☐ From an evolutionary point of view, it has been found that DVA is one of the abilities that **more greatly deteriorates with age**. DVA deterioration is more marked than SVA, and also begins earlier.
  - ☐ A research noted that DVA develops rapidly between 5 and 15 years of age, and that it begins to **decline after the age of 20**.
- **Sports:**
  - ☐ The anticipatory ability based on DVA is crucial to **intercept a moving object** (e.g. a ball) and to **predict the spatial location of items of interest.**
  - ☐ This is the main reason why numerous scientific studies report a **greater DVA for elite athletes compared to sedentary population.**
  - ☐ Moreover, differences have also been found when comparing athletes' DVA in **a dynamic context** (e.g. basketball or tennis) with other modalities with less "visual" requirements such as swimming, **with a marked superiority** in favor of the first.

Beals, R. P., Mayyasi, A. M., Templeton, A. E., &Johnson, W. G. (1971). The relationship between basketball shooting performance and certain visual attributes.

National Research Council's Committee on Vision.(1985). Emergent Techniques for Assessment of Visual Performance.

ACCV 2024

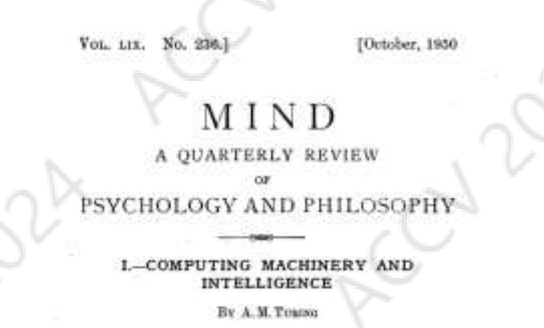
Visual Turing Test

Human tracking abilities are used as a baseline for evaluating machine intelligence.

ACCV 2024

Human-to-Machine Evaluation

- **How to evaluate intelligence? Turing Test**



Vol. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

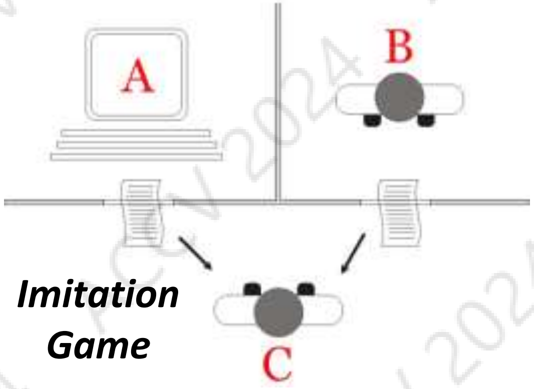
I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

Can machine think?

Computing Machinery and Intelligence. Mind, 1950 (236): 433–460

1950: Alan Turing, the father of artificial intelligence, proposed the **Turing Test**.



Imitation Game

The explain of the Turing test:

- Player C is given the task of trying to determine **which player – A or B – is a computer and which is a human**.
- The player C is limited to using the responses to **written questions** to make the judgment.

➢ The Turing test gives a **concrete and operational way** to measure intelligence and provides **an objective standard** for judging intelligence.

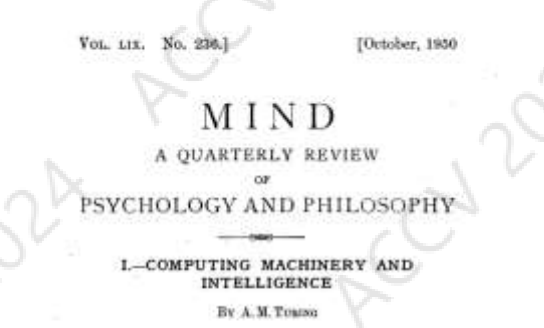
➢ It **avoids unnecessary debates about the nature of intelligence**.

ACCV 2024

Human-to-Machine Evaluation

Introduce human factor is important!

- **How to evaluate intelligence? Turing Test**



Vol. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

Can machine think?

Computing Machinery and Intelligence. Mind, 1950 (236): 433–460

1950: Alan Turing, the father of artificial intelligence, proposed the **Turing Test**.

Keypoint: **Human-Machine Comparison**

Milestone works in decision-making tasks:

Garry Kasparov
VS

1997: DeepBlue defeated Garry Kasparov in international chess competitions.

2016: AlphaGo defeated Lee Sedol in a Go competition.

DeepStack
EXTRAVAGANZA

2017: DeepStack defeated human professional players in Texas Hold'em poker.

ACCV 2024

Human-to-Machine Evaluation

- **How to evaluate visual intelligence? Visual Turing Test.**

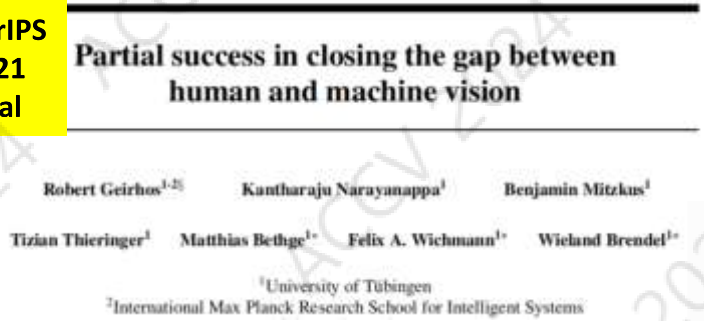
  - **Visual Turing Test** is an evaluation paradigm inspired by the traditional Turing Test, designed to assess whether computer vision systems **possess human-level visual understanding.**
    - The core objective is to compare machine and human performance on visual tasks, determining if the algorithm exhibits sufficient intelligence to **match or exceed human capabilities in complex scenarios.**
  - **Principles of Visual Turing Test:** The Visual Turing Test requires a machine to produce results in **specific visual tasks**, which are then compared to **human results**. Typical tasks include object recognition, tracking, and image classification.
    - In the test, if the machine's performance is highly similar to human results or indistinguishable, it is considered that the machine has achieved human-like visual understanding.

ACCV 2024

Human-to-Machine Evaluation

- **Example 1. Visual Turing Test in Classification**

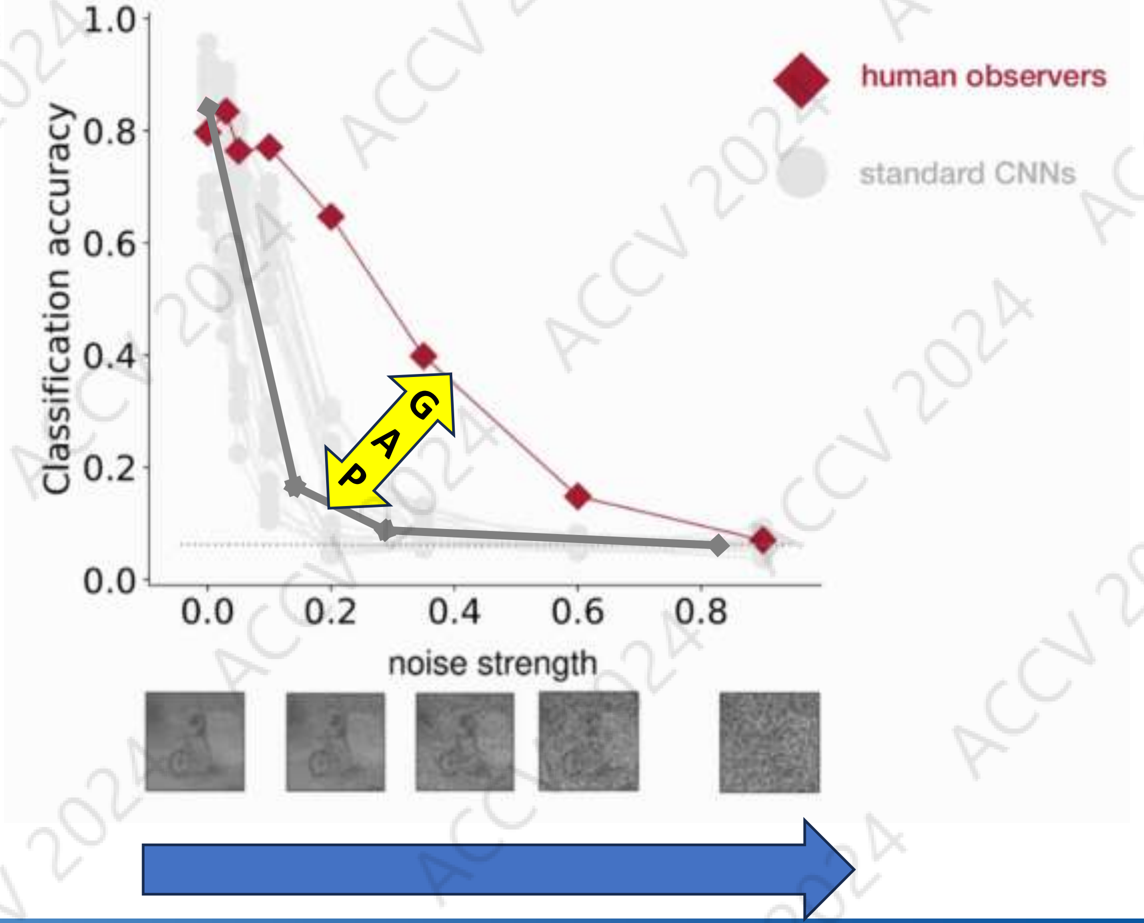
NeurIPS 2021 Oral



Partial success in closing the gap between human and machine vision

Robert Geirhos[1,2§] Kantharaju Narayanappa[1] Benjamin Mitzkus[1]
Tizian Thieringer[1] Matthias Bethge[1*] Felix A. Wichmann[1*] Wieland Brendel[1*]

[1]University of Tübingen
[2]International Max Planck Research School for Intelligent Systems

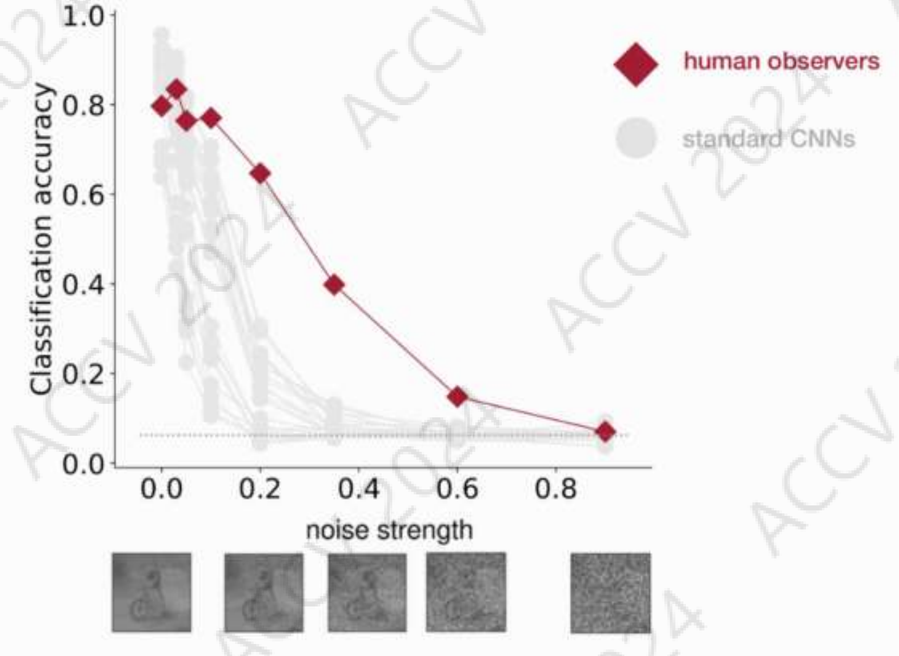


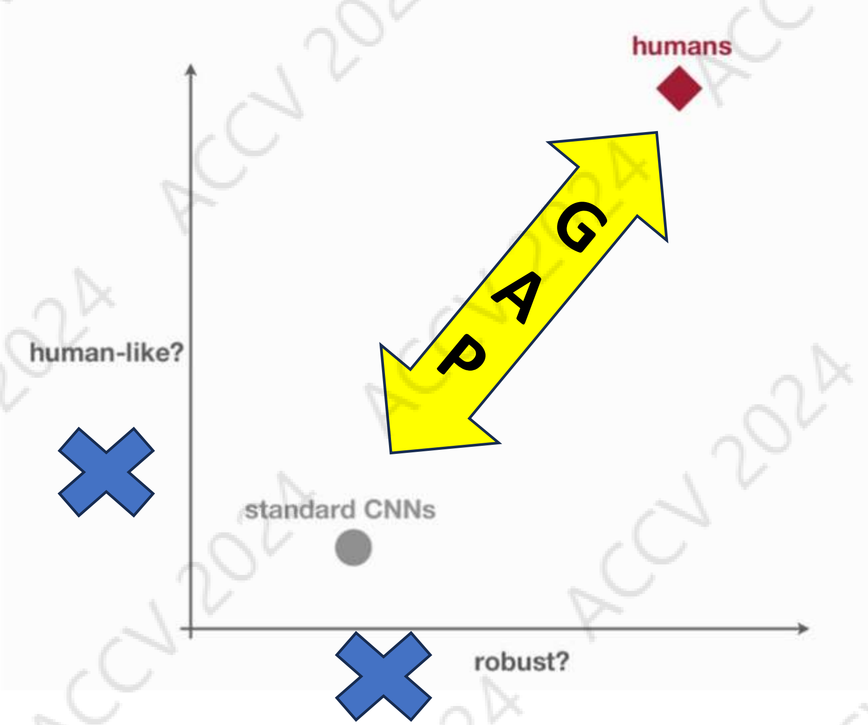
ACCV 2024

Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 23885-23899.

Human-to-Machine Evaluation

- **Example 1. Visual Turing Test in Classification**




humans
G
A
P
human-like?
standard CNNs
robust?

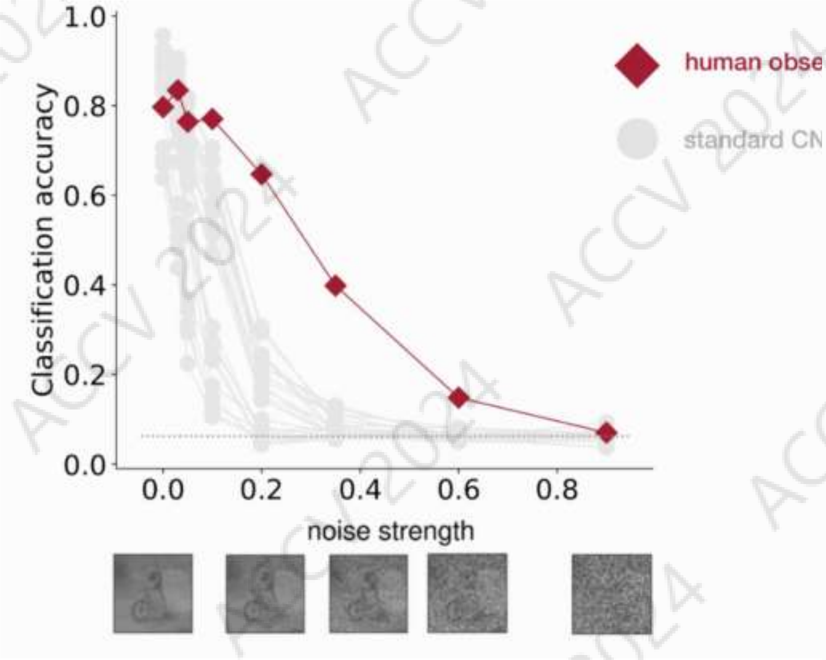
Big robustness gap between CNNs & humans

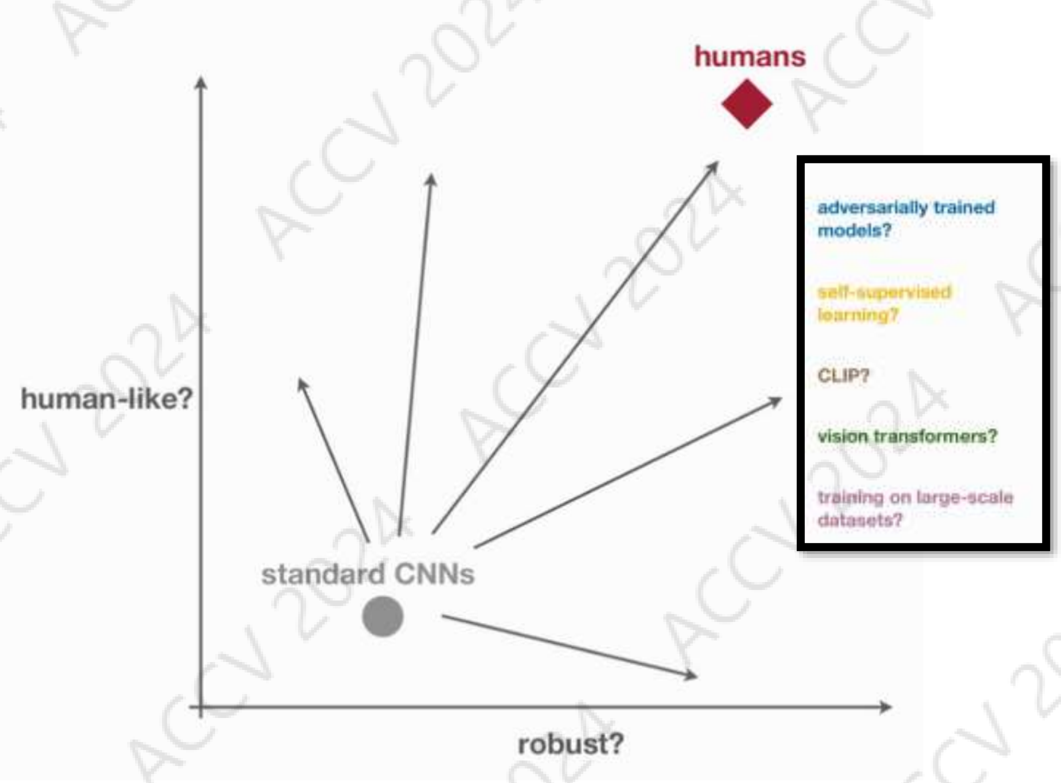
Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 23885-23899.

ACCV 2024

Human-to-Machine Evaluation

- **Example 1. Visual Turing Test in Classification**




humans
human-like?
standard CNNs
robust?
adversarially trained models?
self-supervised learning?
CLIP?
vision transformers?
training on large-scale datasets?

Are we making progress in closing the gap between human and machine vision?

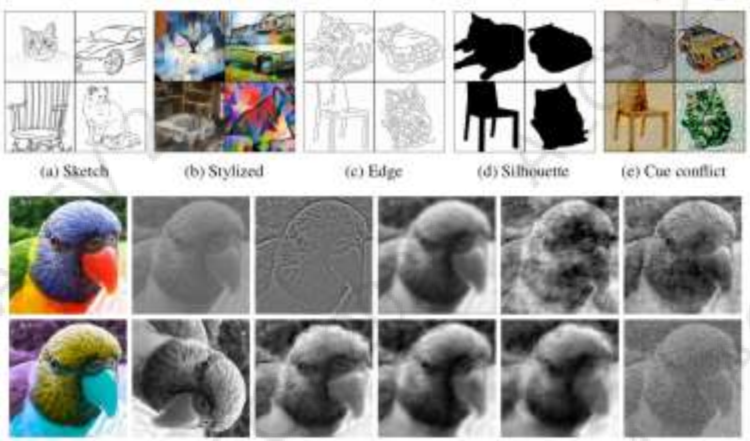
Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 23885-23899.

ACCV 2024

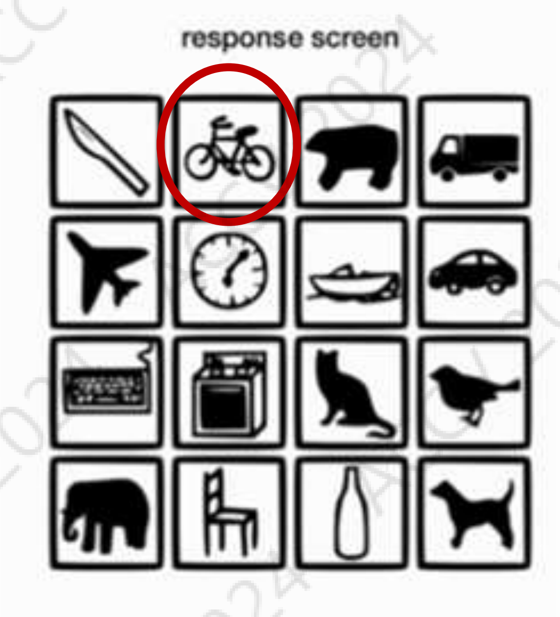
Human-to-Machine Evaluation

- **Example 1. Visual Turing Test in Classification**

Are we making progress in closing the gap between human and machine vision?


(a) Sketch
(b) Stylized
(c) Edge
(d) Silhouette
(e) Cue conflict

17 image processing styles


response screen

Participants were explained how to respond (via **mouse click**), instructed to respond **as accurately as possible**, and to go with their **best guess if unsure**.

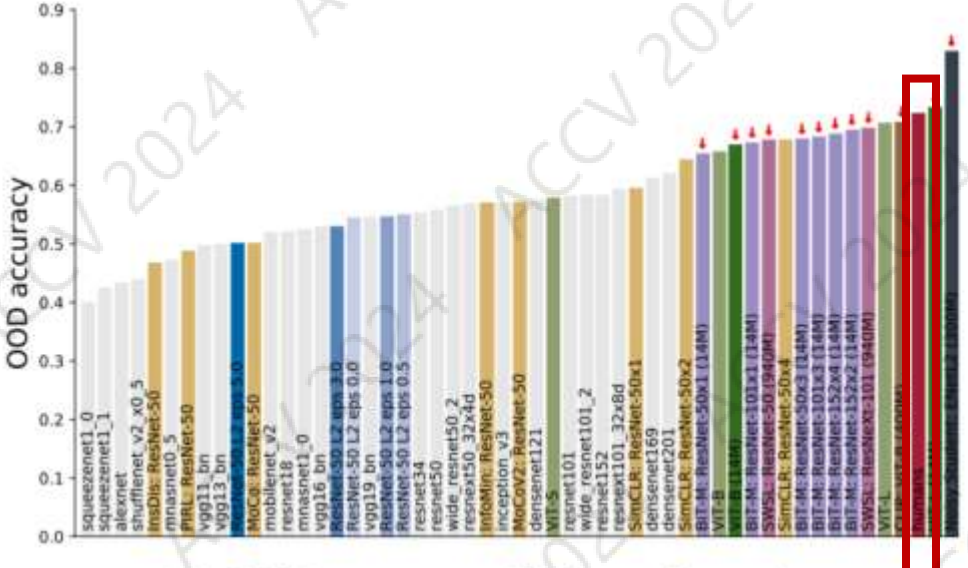
Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 23885-23899.

ACCV 2024

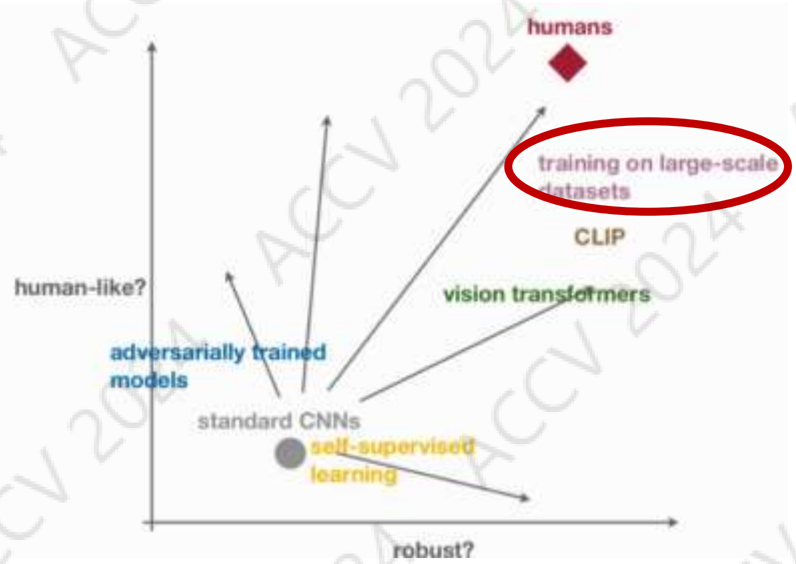
Human-to-Machine Evaluation

- **Example 1. Visual Turing Test in Classification**

Are we making progress in closing the gap between human and machine vision?



(a) OOD accuracy (higher = better).


humans
training on large-scale datasets
CLIP
vision transformers
human-like?
adversarially trained models
standard CNNs
self-supervised learning
robust?

The longstanding OOD robustness gap between human and machine vision is closing.

Geirhos R, Narayanappa K, Mitzkus B, et al. Partial success in closing the gap between human and machine vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 23885-23899.

Human-to-Machine Evaluation

- **Example 2. Visual Turing Test in Image Distortion**

Patterns

CellPress OPEN ACCESS

Cell Patterns 2023

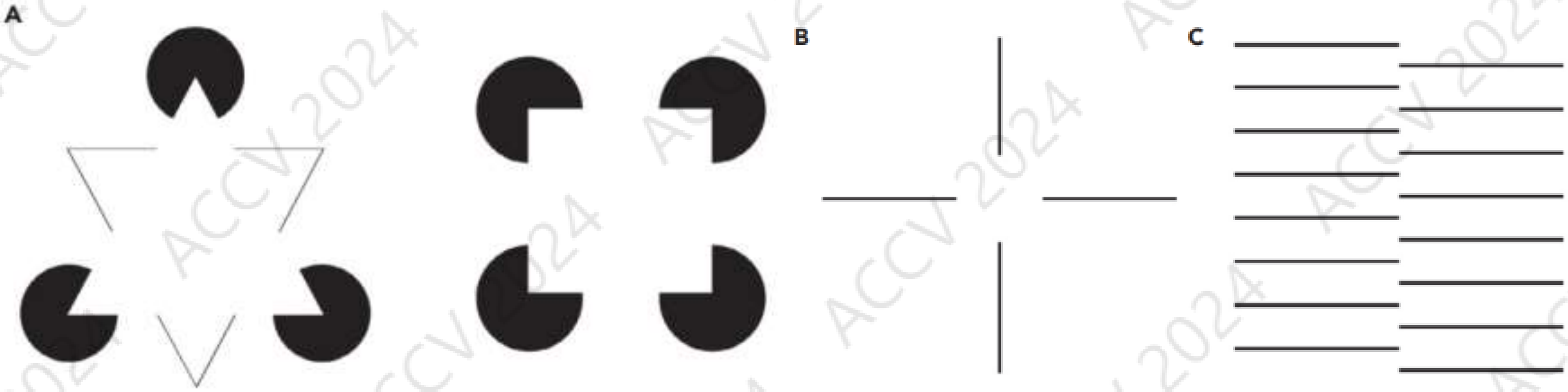
Article

Challenging deep learning models with image distortion based on the abutting grating illusion

Jinyu Fan[1,6] and Yi Zeng[1,2,3,4,5,6,7,*]

[1]Brain-inspired Cognitive Intelligence Lab, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[5]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China
[6]These authors contributed equally
[7]Lead contact
*Correspondence: yi.zeng@ia.ac.cn
https://doi.org/10.1016/j.patter.2023.100695

Illusory contours **evoke the** **perception** **of a distinct boundary** **without color contrast or luminance gradients** **across that boundary.**



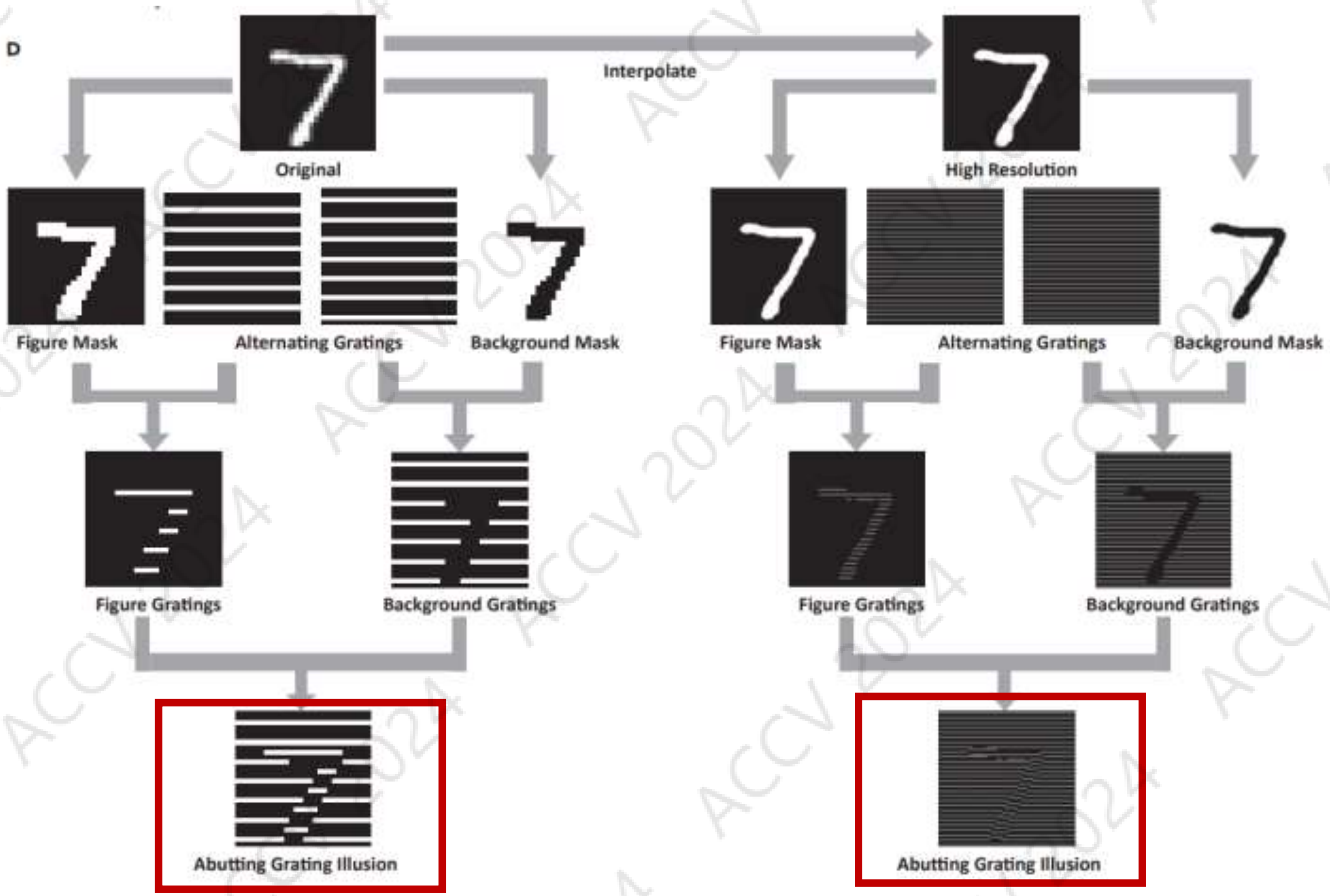
Kanizsa triangle and Kanizsa square **Ehrenstein illusion** **Abutting grating illusion**

Fan J, Zeng Y. Challenging deep learning models with image distortion based on the abutting grating illusion[J]. Patterns, 2023, 4(3).

ACCV 2024

Human-to-Machine Evaluation

- **Example 2. Visual Turing Test in Image Distortion**


D
Original
Interpolate
High Resolution
Figure Mask
Alternating Gratings
Background Mask
Figure Mask
Alternating Gratings
Background Mask
Figure Gratings
Background Gratings
Figure Gratings
Background Gratings
Abutting Grating Illusion
Abutting Grating Illusion

AG-MNIST (1*28*28)

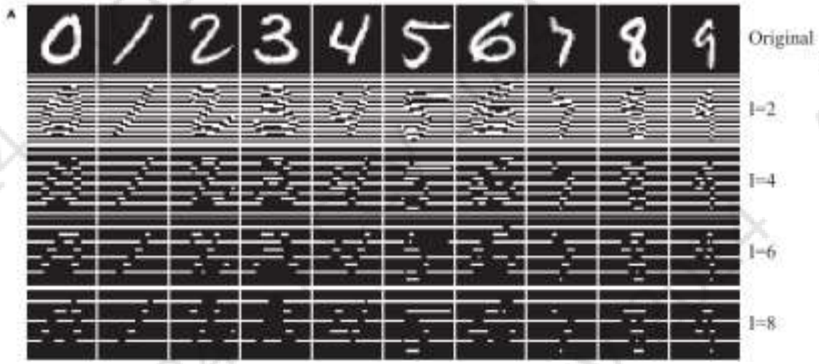
High-resolution AG-MNIST (3*224*224)

Fan J, Zeng Y. Challenging deep learning models with image distortion based on the abutting grating illusion[J]. Patterns, 2023, 4(3).

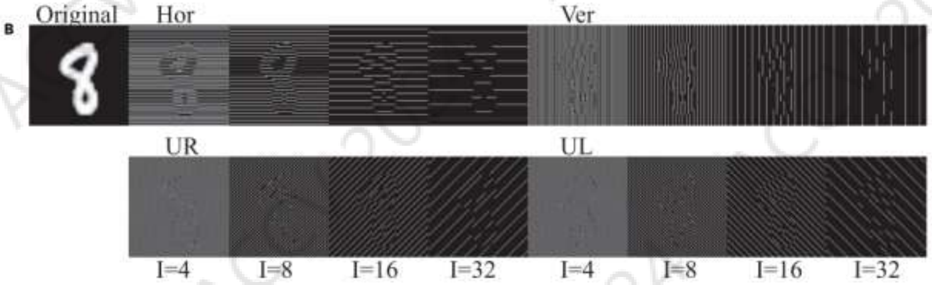
ACCV 2024

Human-to-Machine Evaluation

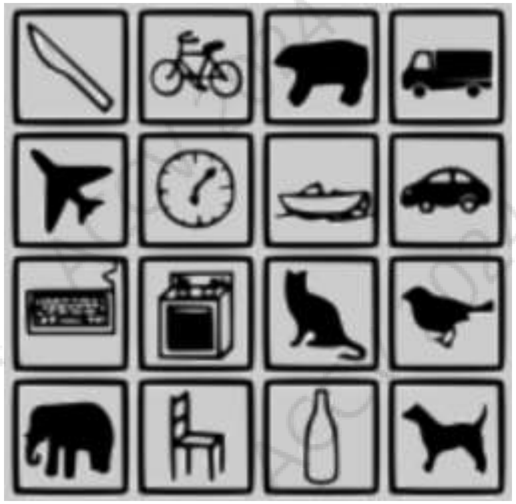
- **Example 2. Visual Turing Test in Image Distortion**


A
Original
I=2
I=4
I=6
I=8

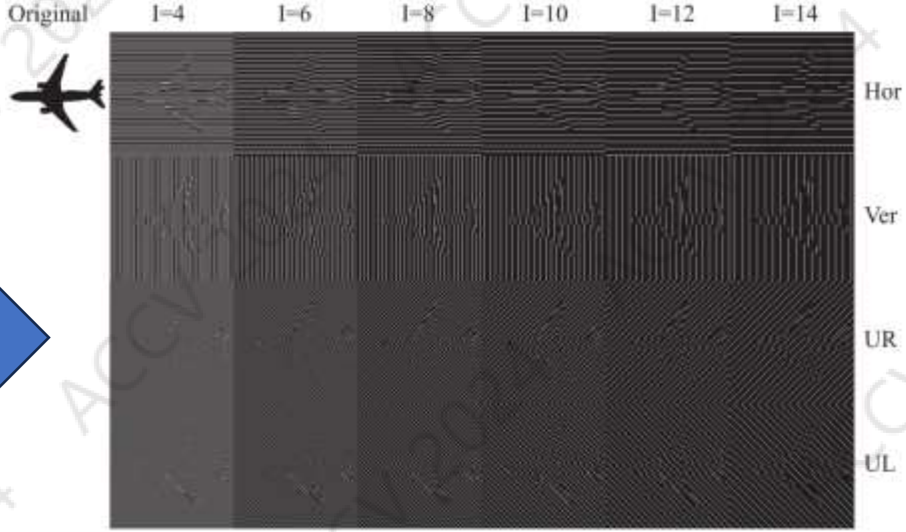
AG-MNIST:
Different spacing


B
Original
Hor
Ver
UR
UL
I=4
I=8
I=16
I=32
I=4
I=8
I=16
I=32

High-resolution AG-MNIST:
Different spacing & Different orientations



Silhouettes


Original
I=4
I=6
I=8
I=10
I=12
I=14
Hor
Ver
UR
UL

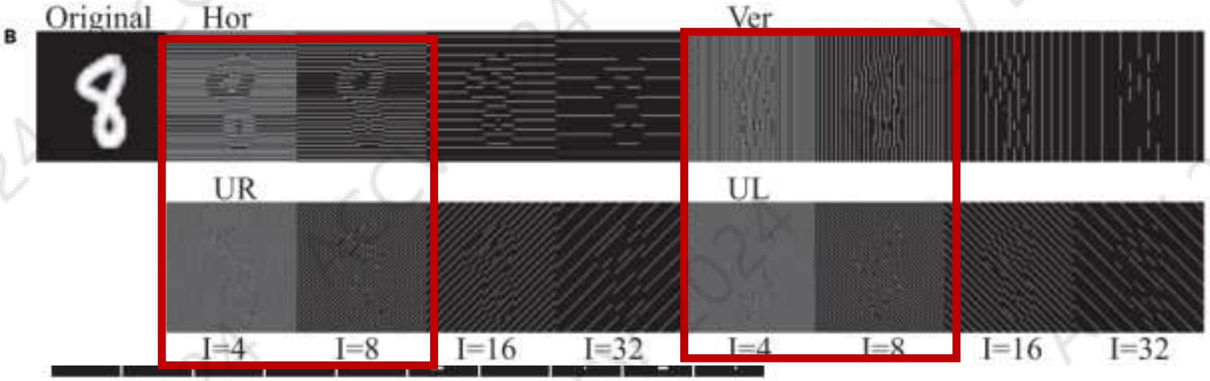
AG-silhouettes:
Different spacing & Different orientations

Fan J, Zeng Y. Challenging deep learning models with image distortion based on the abutting grating illusion[J]. Patterns, 2023, 4(3).

ACCV 2024

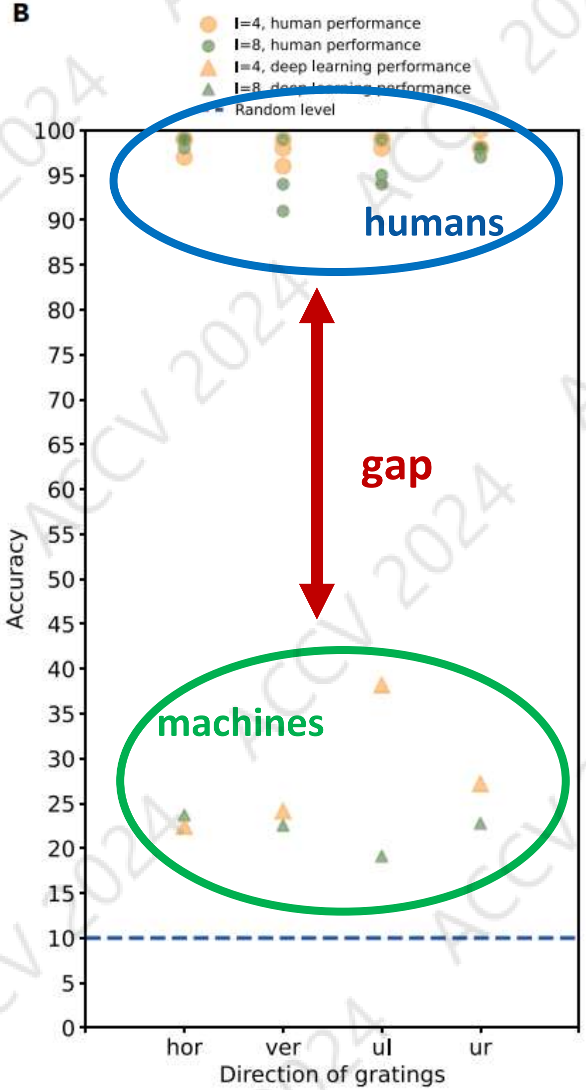
Human-to-Machine Evaluation

- **Example 2. Visual Turing Test in Image Distortion**


B
Original
Hor
Ver
UR
UL
I=4
I=8
I=16
I=32
I=4
I=8
I=16
I=32

High-resolution AG-MNIST

The gap between humans and deep learning models is still immense.

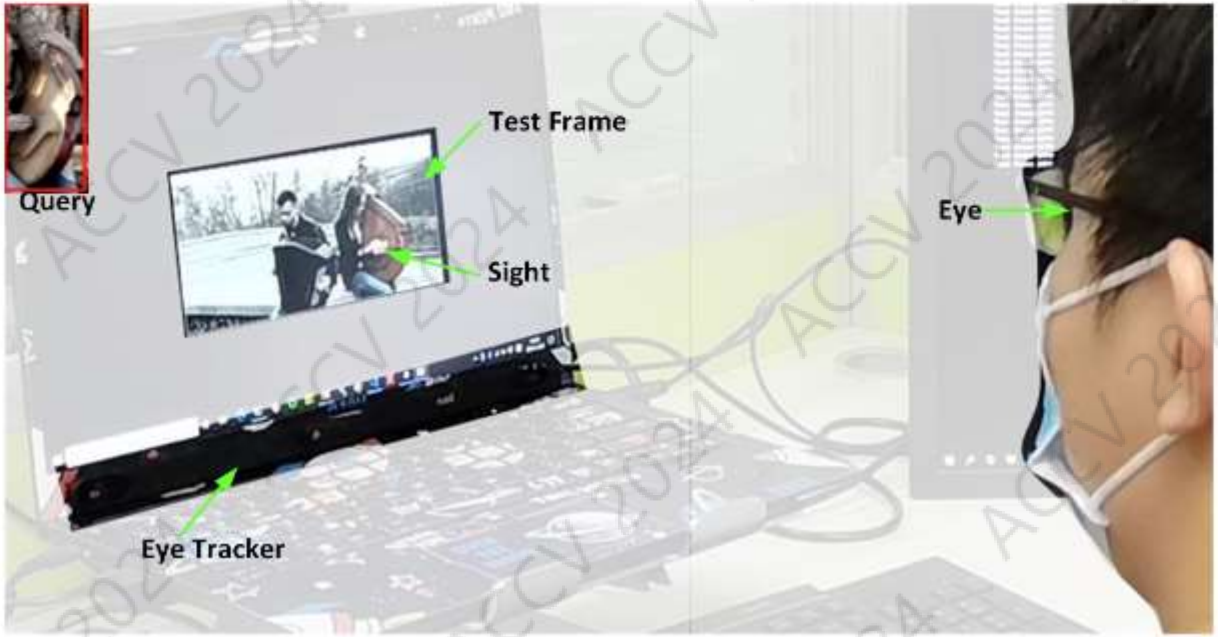


Fan J, Zeng Y. Challenging deep learning models with image distortion based on the abutting grating illusion[J]. Patterns, 2023, 4(3).

ACCV 2024

Human-to-Machine Evaluation

- **Example 3. Visual Turing Test in Global Instance Tracking**



Schematic diagram of the human visual tracking experiment.

- **Eye-Tracking Experiments:** By using **eye-tracking devices**, human visual tracking data, such as gaze focus points and eye movements during a tracking task, are collected. This data reflects how humans track objects in visual tasks and serves as a reference for evaluating the performance of machine vision.
  - **For the first time**, human participants are introduced into the evaluation process of single object tracking tasks.

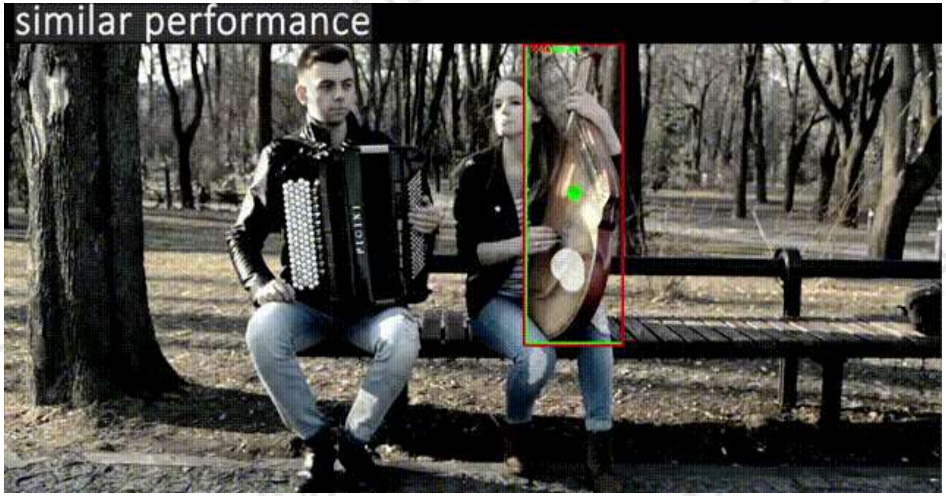
*S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

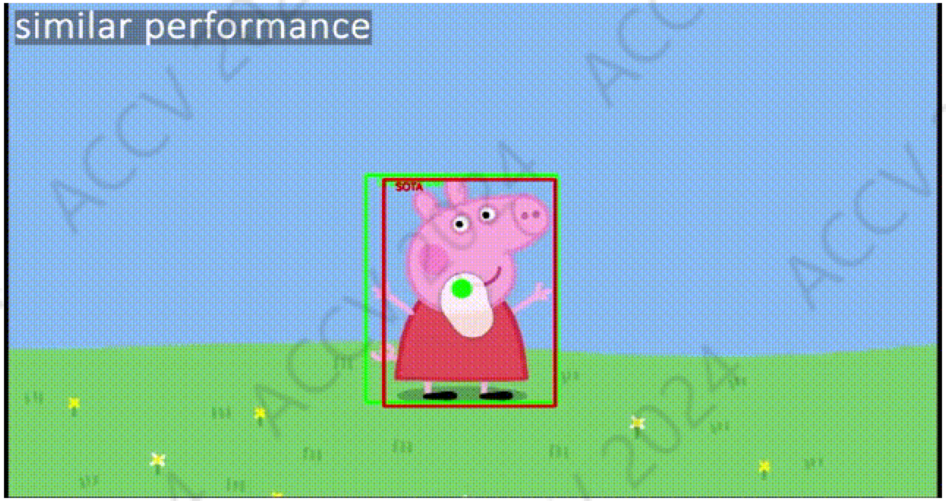
ACCV 2024

Human-to-Machine Evaluation

- **Example 3. Visual Turing Test in Global Instance Tracking**

➢ **When the target moves smoothly:** Humans and machines perform similarly





- **White semi-transparent dot:** human eye tracking result
- **Green rectangle:** target position
- **Green dot:** target center
- **Red rectangle:** SOTA algorithm tracking result

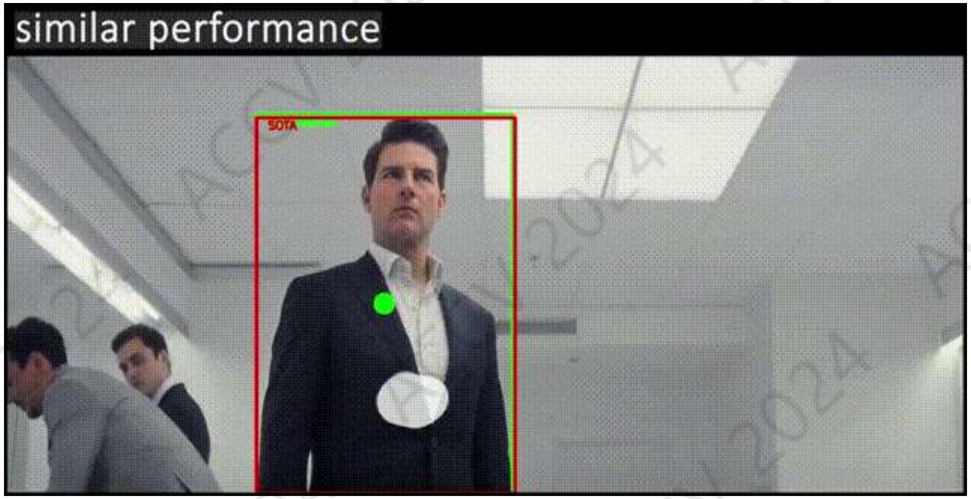
ACCV 2024

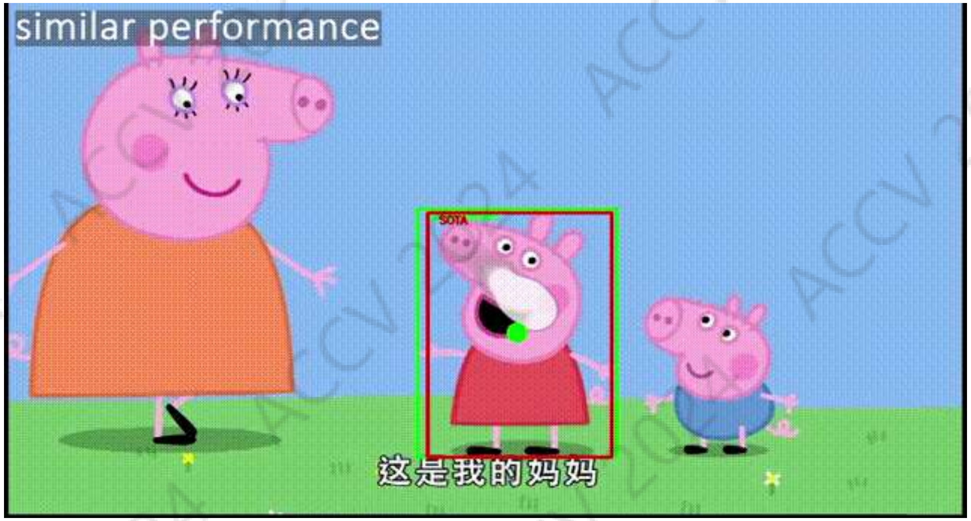
S. Hu, X. Zhao#, *L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

Human-to-Machine Evaluation

- **Example 3. Visual Turing Test in Global Instance Tracking**

➢ **A few challenging factors:** Humans are better than machines


similar performance
SOTA


similar performance
SOTA
这是我的妈妈

- **White semi-transparent dot:** human eye tracking result
- **Green rectangle:** target position
- **Green dot:** target center
- **Red rectangle:** SOTA algorithm tracking result

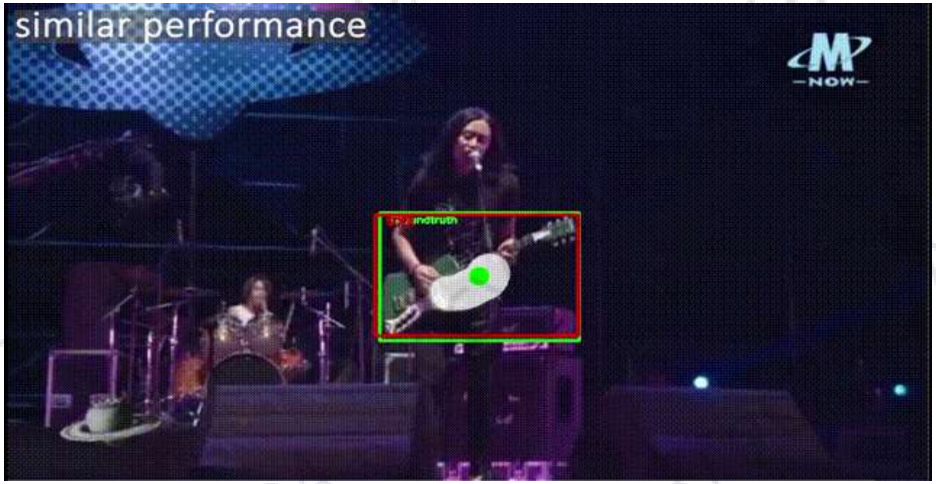
__S. Hu, X. Zhao#__, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (__TPAMI__), vol. 45, no. 1, pp. 576–592, 2023.

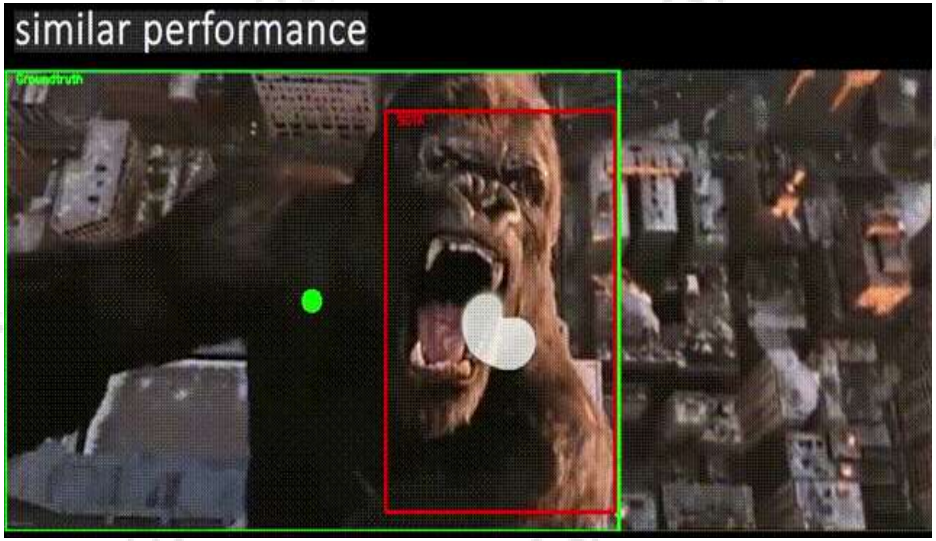
ACCV 2024

Human-to-Machine Evaluation

- **Example 3. Visual Turing Test in Global Instance Tracking**

➢ **Multiple challenging factors:** Both failed


similar performance
M NOW
Groundtruth


similar performance
Groundtruth
SOTA

- **White semi-transparent dot:** human eye tracking result
- **Green rectangle:** target position
- **Green dot:** target center
- **Red rectangle:** SOTA algorithm tracking result

*S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

ACCV 2024

Human-to-Machine Evaluation

- ## Core Steps in Visual Turing Test

  - **Task Design:** Design identical visual tasks for both the machine and human participants.
  - **Result Collection:** Collect the results from both the machine and human participants on the same tasks.
  - **Result Comparison:** Use evaluation metrics such as similarity measures or success rates to compare machine and human performance.
  - **Judgment:** If the machine’s results are indistinguishable from the human results, the machine is considered to have passed the Visual Turing Test.

ACCV 2024

Conclusion

- **Humans:**
  - **Visual Theories**
    - Feature Integration Theory
    - Recognition-by-Components
    - Visual Computation
  - **Visual Abilities**
    - Static Visual Ability
    - Dynamic Visual Ability
  - **Insights for AI Development:** Understanding human visual theories and abilities, **provides foundational insights for enhancing algorithms**. By emulating these human mechanisms, machine vision systems can achieve **more accurate and adaptable tracking performance.**

ACCV 2024

Conclusion

- **Human-to-Machine Evaluation:**
  - This approach **brings in human factors**, such as perception of occlusions, complex background differentiation, and rapid adjustments to changing conditions.
  - By integrating human perspectives, human-to-machine evaluation enables **a more holistic understanding of a tracker's capabilities**, moving beyond traditional metrics to capture qualitative aspects of **intelligence and decision-making**.

ACCV 2024

CONTENTS

- Introduction
- Part 1. Visual Object Tracking Task
- Part 2. Experimental Environments
- Part 3. Algorithms and Traditional Machine-Machine Comparisons
- Part 4. Human Visual Abilities and Visual Turing Test
- **Trends and Future Directions**
    - **More Human-like Task Design**
    - **More Realistic Data Environment**
    - **More Human-like Executors**
    - **More Intelligent Evaluation**

Scan to download this tutorial PPT

ACCV 2024

Trend 1. More Human-like Task Design

ACCV 2024

Trend 1. More Human-like Task Design

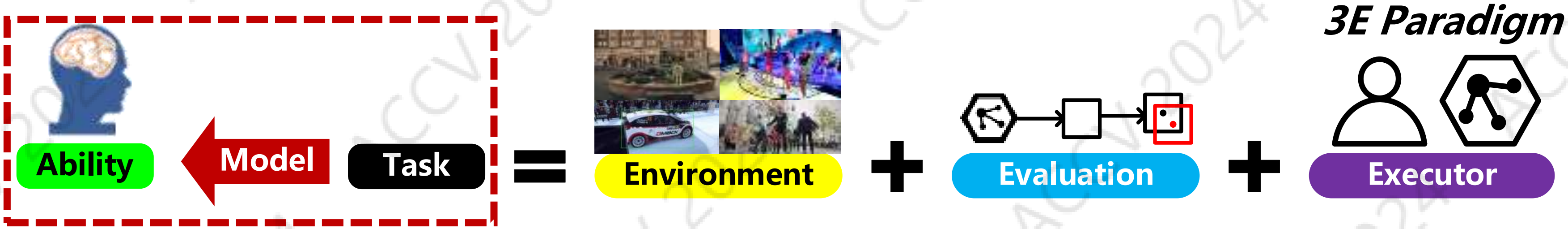
- **What are the abilities of humans? → Designing more human-like task to model the dynamic vision ability.**

Short-term tracking & Long-term tracking:
Methods utilize local search to locate the target near to its position in the previous frame.

→ **perceptual level ability**

Shotcut Start: False
Occlusion: True

Global Instance Tracking:
Methods should remember the target and re-detect it in a new shot.
→ **cognitive level ability**



3E Paradigm

Ability ← **Model** **Task** = **Environment** + **Evaluation** + **Executor**

S. Hu, X. Zhao#, L. Huang, et al., *"Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), vol. 45, no. 1, pp. 576–592, 2023.*

ACCV 2024

Trend 1. More Human-like Task Design

- **What are the abilities of humans? → Designing more human-like task to model the dynamic vision ability.**

Short-term tracking & Long-term tracking:
Methods utilize local search to locate the target near to its position in the previous frame.

→ **perceptual level ability**

Occlusion: True

Global Instance Tracking:
Methods should remember the target and re-detect it in a new shot.
→ **cognitive level ability**

Visual Information (VOT)

→

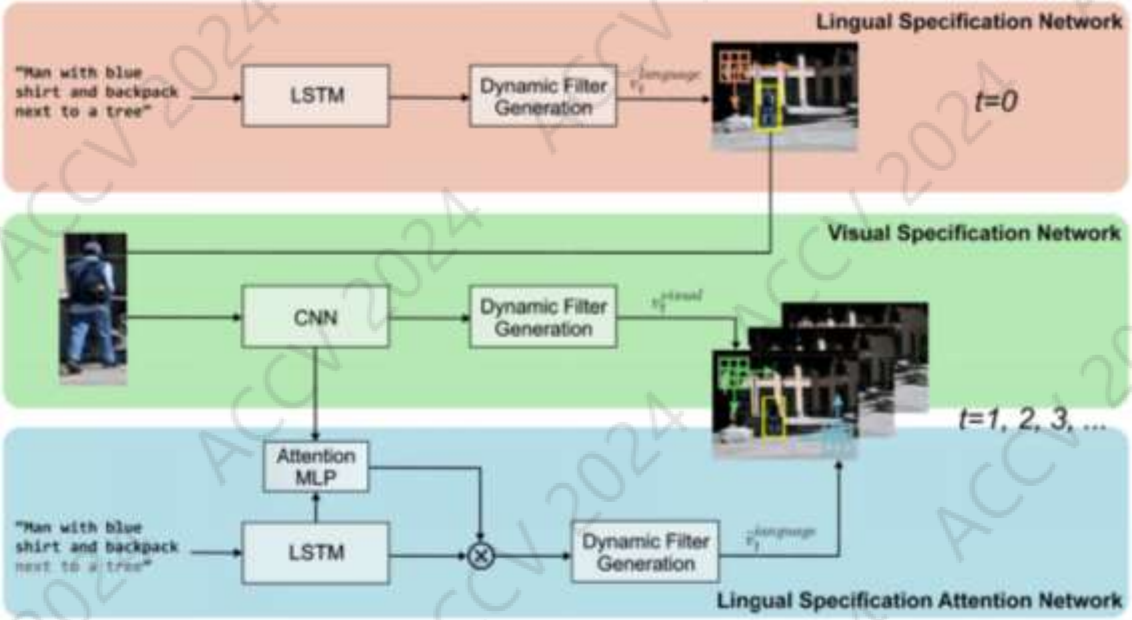
Add Semantic Information (VLT)

S. Hu, X. Zhao#, L. Huang, et al., "Global instance tracking: Locating target more like humans," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 45, no. 1, pp. 576–592, 2023.

ACCV 2024

Trend 1. More Human-like Task Design

- **Visual Language Tracking**


Lingual Specification Network
"Man with blue shirt and backpack next to a tree"
LSTM
Dynamic Filter Generation
$v_t^{language}$
t=0
Visual Specification Network
CNN
Dynamic Filter Generation
v_t^{visual}
t=1, 2, 3, ...
Attention MLP
"Man with blue shirt and backpack next to a tree"
LSTM
⊗
Dynamic Filter Generation
$v_t^{language}$
Lingual Specification Attention Network

- **Type 1:** Tracking **using only textual information (Grounding)** without relying on visual data. The target is located and tracked purely through natural language descriptions.
- **Type 2:** The target is **initially located using textual descriptions**, and **then visual data is used** for single object tracking without further input from text.
- **Type 3:** **Both text and bounding boxes are used** for target initialization, and text continues to aid in object tracking throughout the process.

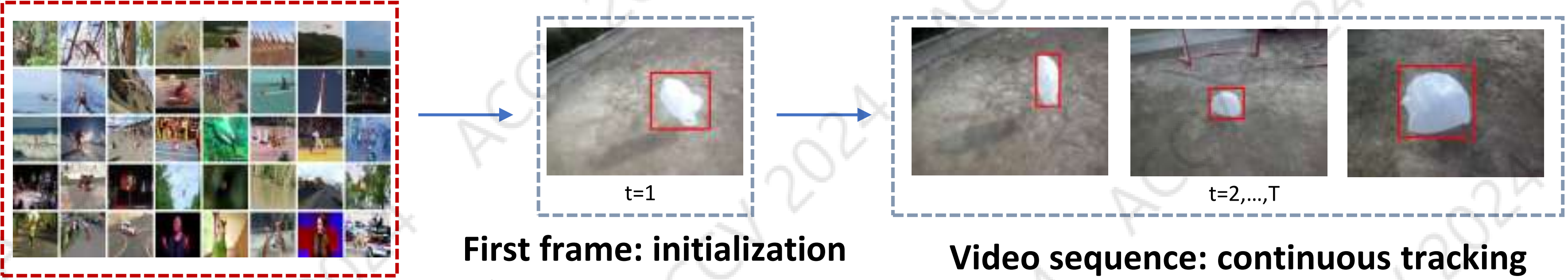
Li Z, Tao R, Gavves E, et al. Tracking by natural language specification, CVPR 2017

ACCV 2024

Trend 1. More Human-like Task Design

- **Visual Language Tracking**

➢ **Deep Learning-based Tracking Methods:** The **majority of deep learning-based object tracking methods belong to Type 3**, where both text and bounding box (BBox) are used for initialization, and text may or may not be used for further tracking. A **smaller portion of methods combine Type 2**, where text is only used during initialization, and further tracking relies solely on visual information.



Pre-training dataset

First frame: initialization
(semantic information and/or Bbox)

Video sequence: continuous tracking
(Use or not use semantic information)

OTB99-Lang (2017)

Bear-12: "white bear walking on grass around the river bank"
Bus-19: "red bus running on the highway"
Horse-1: "brown horse running on the ground"

LaSOT (2019&2021)

TNL2K (2021)

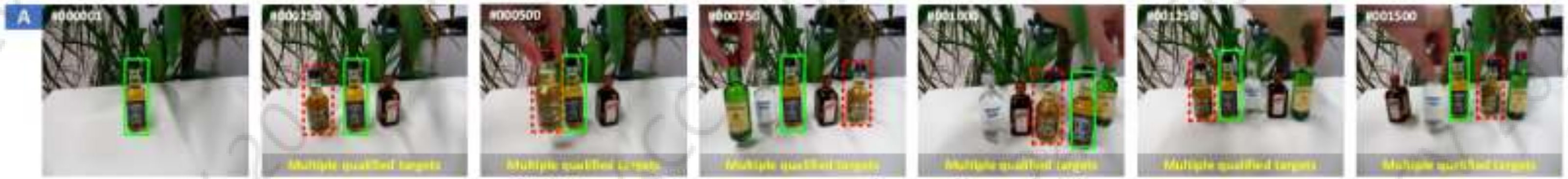
MGIT (2023)

Representative experimental environment

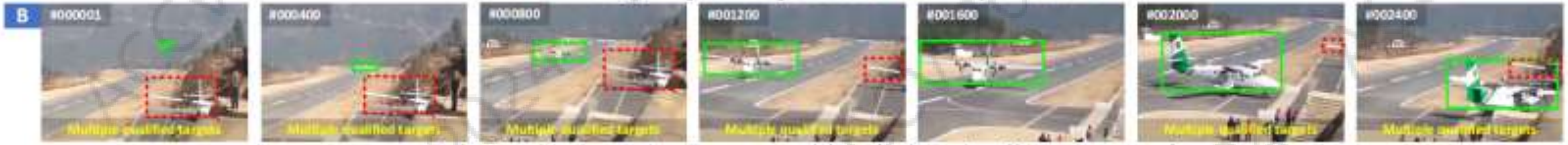
ACCV 2024

Trend 1. More Human-like Task Design

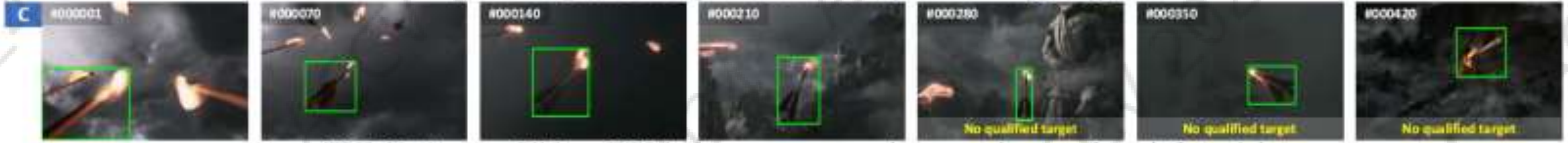
- **Example: Multimodal Global Instance Tracking (MGIT)**
  - ➢ **Motivation:** The VLT / VOT algorithm **performs poorly in complex scenes (long sequences & complex spatio-temporal causal relationships).** Some recent researches have considered studying from a multi-modal perspective:
    - **Limitations 1. Short sequence** (from hundreds of frames to thousands of frames) **→ Simple narrative content**
    - **Limitations 2. Inaccurate semantic annotation** (describing only the information of the first frame, and there may be multiple objects in the scene that fit the description) **→ Misguide algorithms**


A
#000001
#000250
#000500
#000750
#001000
#001250
#001500
Multiple qualified targets

OTB-Lang Liquor sequence: brown liquor bottle


B
#000001
#000400
#000800
#001200
#001600
#002000
#002400
Multiple qualified targets

LaSOT airplane-1 sequence: white airplane landing on ground


C
#000001
#000070
#000140
#000210
#000280
#000350
#000420
No qualified target

TNL2k Arrow_Video_ZZ04_done sequence: the second arrow from left to right

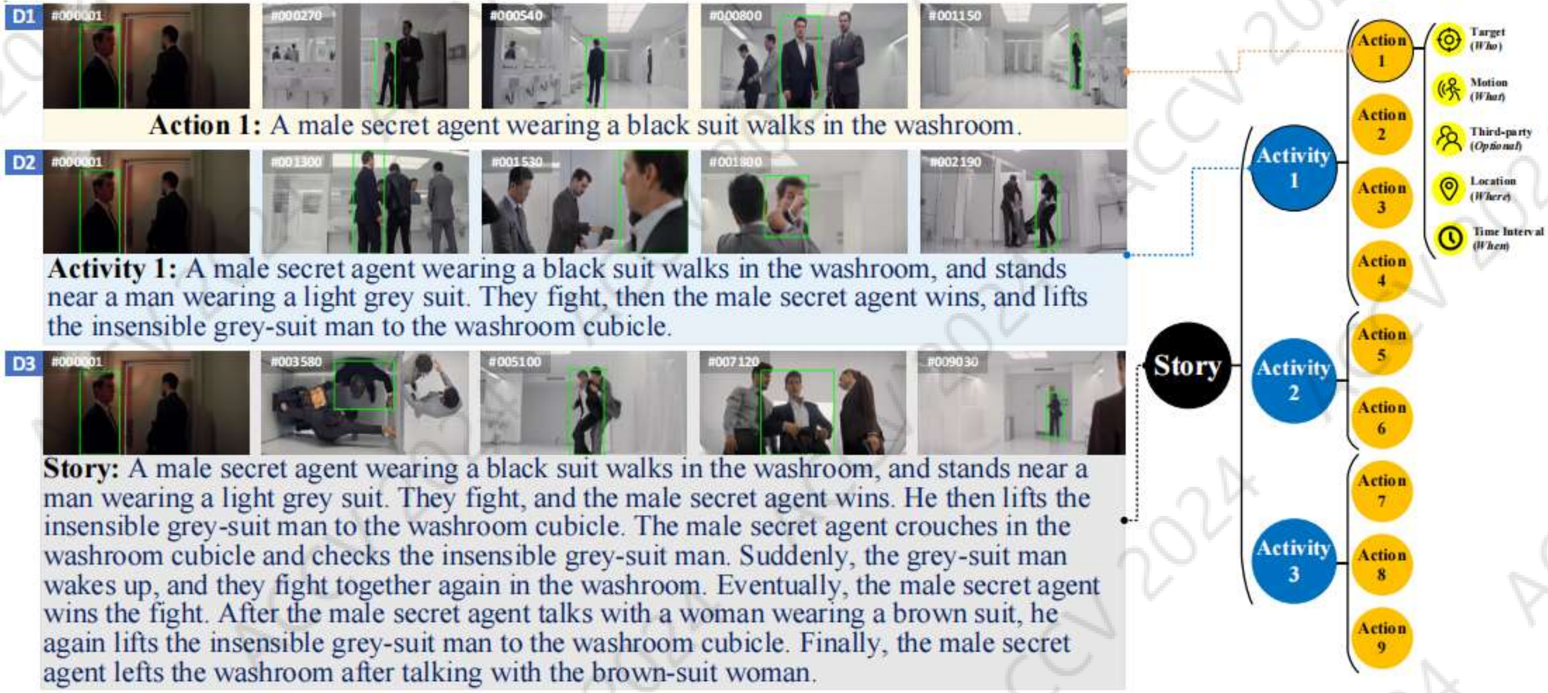
Limitations of existing works

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

ACCV 2024

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**
  - **Limitations 1. Short sequence** (from hundreds of frames to thousands of frames) → **Simple narrative content** → **Using longer sequences with more complex narratives**
  - **Limitations 2. Inaccurate semantic annotation** (describing only the information of the first frame, and there may be multiple objects in the scene that fit the description) → **Misguide algorithms** → **Design a multi-granular annotation strategy to portray long videos**


D1
#000001
#000270
#000540
#000800
#001150
D2
#000001
#001300
#001530
#001800
#002190
D3
#000001
#003580
#005100
#007120
#009030
Story
Activity 1
Activity 2
Activity 3
Action 1
Action 2
Action 3
Action 4
Action 5
Action 6
Action 7
Action 8
Action 9
Target (Who)
Motion (What)
Third-party (Optional)
Location (Where)
Time Interval (When)

Action 1: A male secret agent wearing a black suit walks in the washroom.

Activity 1: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, then the male secret agent wins, and lifts the insensible grey-suit man to the washroom cubicle.

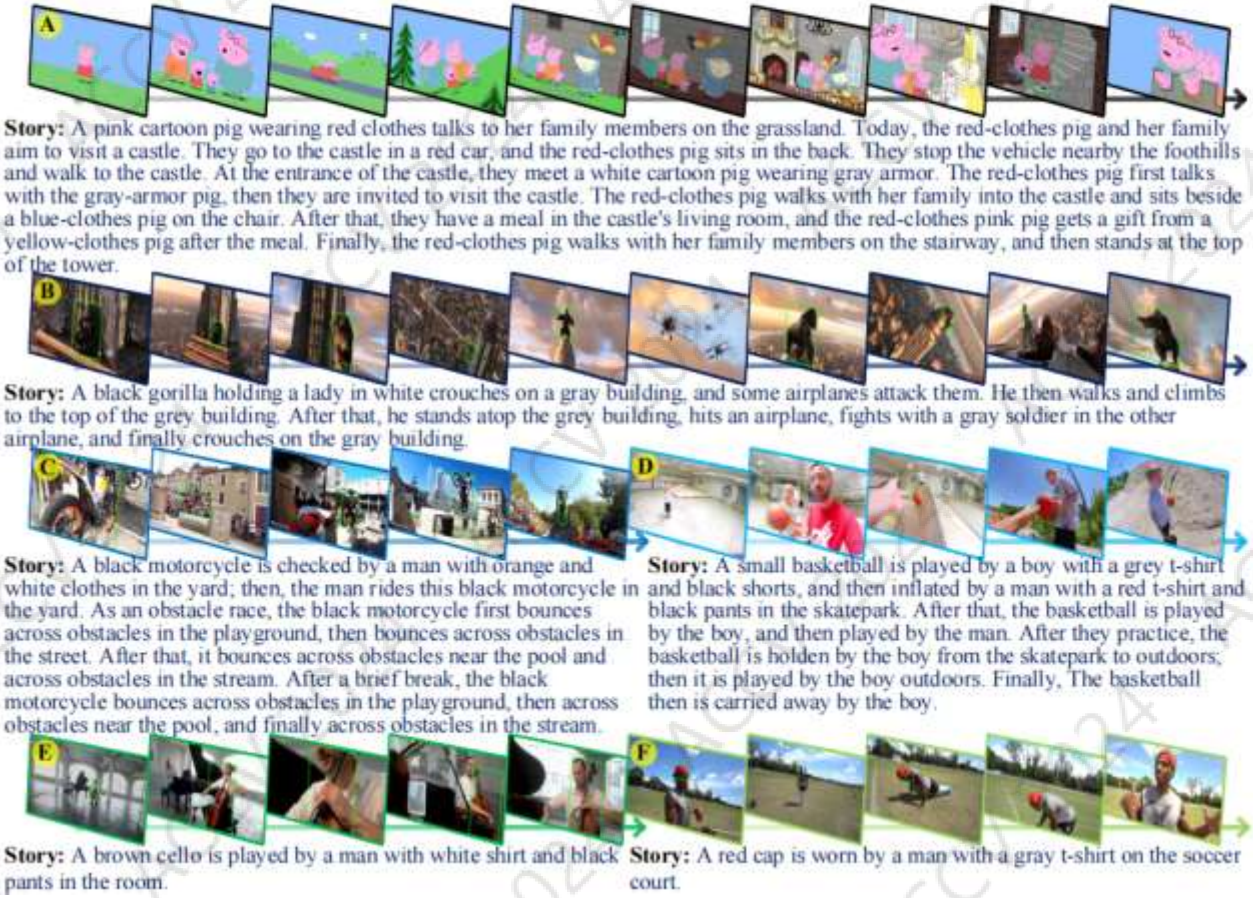
Story: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, and the male secret agent wins. He then lifts the insensible grey-suit man to the washroom cubicle. The male secret agent crouches in the washroom cubicle and checks the insensible grey-suit man. Suddenly, the grey-suit man wakes up, and they fight together again in the washroom. Eventually, the male secret agent wins the fight. After the male secret agent talks with a woman wearing a brown suit, he again lifts the insensible grey-suit man to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

Structure of MGIT

S. Hu, D. Zhang, M. Wu, et al., *“A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship,” in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**
  - **Rich subject matter →** Sufficiently covers the complex spatio-temporal causal relationships of long videos



Story: A pink cartoon pig wearing red clothes talks to her family members on the grassland. Today, the red-clothes pig and her family aim to visit a castle. They go to the castle in a red car, and the red-clothes pig sits in the back. They stop the vehicle nearby the foothills and walk to the castle. At the entrance of the castle, they meet a white cartoon pig wearing gray armor. The red-clothes pig first talks with the gray-armor pig, then they are invited to visit the castle. The red-clothes pig walks with her family into the castle and sits beside a blue-clothes pig on the chair. After that, they have a meal in the castle's living room, and the red-clothes pink pig gets a gift from a yellow-clothes pig after the meal. Finally, the red-clothes pig walks with her family members on the stairway, and then stands at the top of the tower.

Story: A black gorilla holding a lady in white crouches on a gray building, and some airplanes attack them. He then walks and climbs to the top of the grey building. After that, he stands atop the grey building, hits an airplane, fights with a gray soldier in the other airplane, and finally crouches on the gray building.

Story: A black motorcycle is checked by a man with orange and white clothes in the yard; then, the man rides this black motorcycle in the yard. As an obstacle race, the black motorcycle first bounces across obstacles in the playground, then bounces across obstacles in the street. After that, it bounces across obstacles near the pool and across obstacles in the stream. After a brief break, the black motorcycle bounces across obstacles in the playground, then across obstacles near the pool, and finally across obstacles in the stream.

Story: A small basketball is played by a boy with a grey t-shirt and black shorts, and then inflated by a man with a red t-shirt and black pants in the skatepark. After that, the basketball is played by the boy, and then played by the man. After they practice, the basketball is holden by the boy from the skatepark to outdoors; then it is played by the boy outdoors. Finally, The basketball then is carried away by the boy.

Story: A brown cello is played by a man with white shirt and black pants in the room.

Story: A red cap is worn by a man with a gray t-shirt on the soccer court.

Several examples of MGIT

ACCV 2024

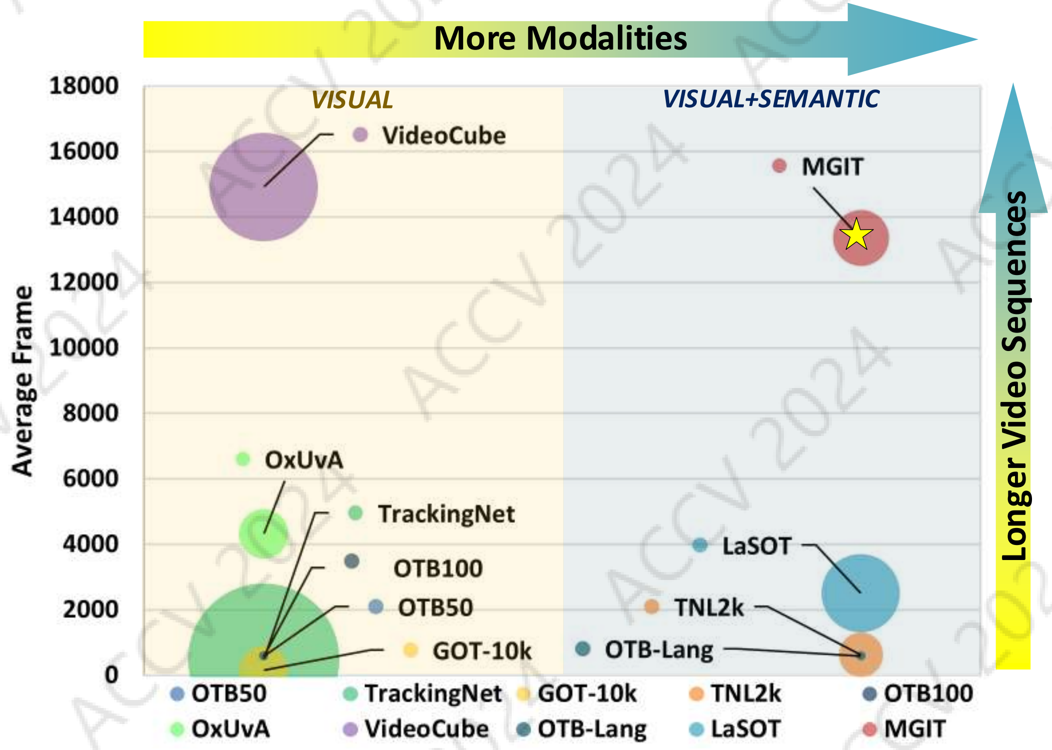
S. Hu, D. Zhang, M. Wu, et al., *"A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**

➢ **Large-scale, Multi-modal**

  - 150 long videos
  - **2.03 million frames**
  - The average length of a single video is **13,500 frames**



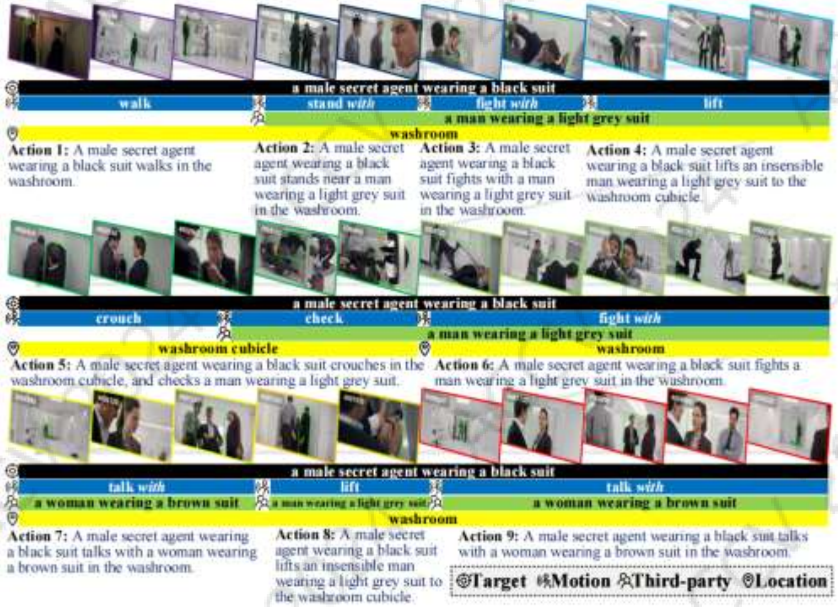
Comparison between MGIT with other VOT / VLT benchmarks

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 1. More Human-like Task Design

• Example: Multimodal Global Instance Tracking (MGIT)

- ➢ Applying **hierarchical structure inspired by human cognition** for multi-granular annotation
  - **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
    - ☐ **Natural Language Grammar Structure :** Subject, Predicate, Object, Adverbial of time, Adverbial of place
    - ☐ **Video Narrative Content :** Time, Location, Character, Event


a male secret agent wearing a black suit
walk
stand *with*
fight *with*
lift
a man wearing a light grey suit
washroom
a male secret agent wearing a black suit
crouch
check
fight *with*
a man wearing a light grey suit
washroom cubicle
washroom
a male secret agent wearing a black suit
talk *with*
lift
talk *with*
a woman wearing a brown suit
a man wearing a light grey suit
a woman wearing a brown suit
washroom
Target
Motion
Third-party
Location

Action 1: A male secret agent wearing a black suit walks in the washroom.

Action 2: A male secret agent wearing a black suit stands near a man wearing a light grey suit in the washroom.

Action 3: A male secret agent wearing a black suit fights with a man wearing a light grey suit in the washroom.

Action 4: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 5: A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit.

Action 6: A male secret agent wearing a black suit fights a man wearing a light grey suit in the washroom.

Action 7: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

Action 8: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 9: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

An example of action annotation

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

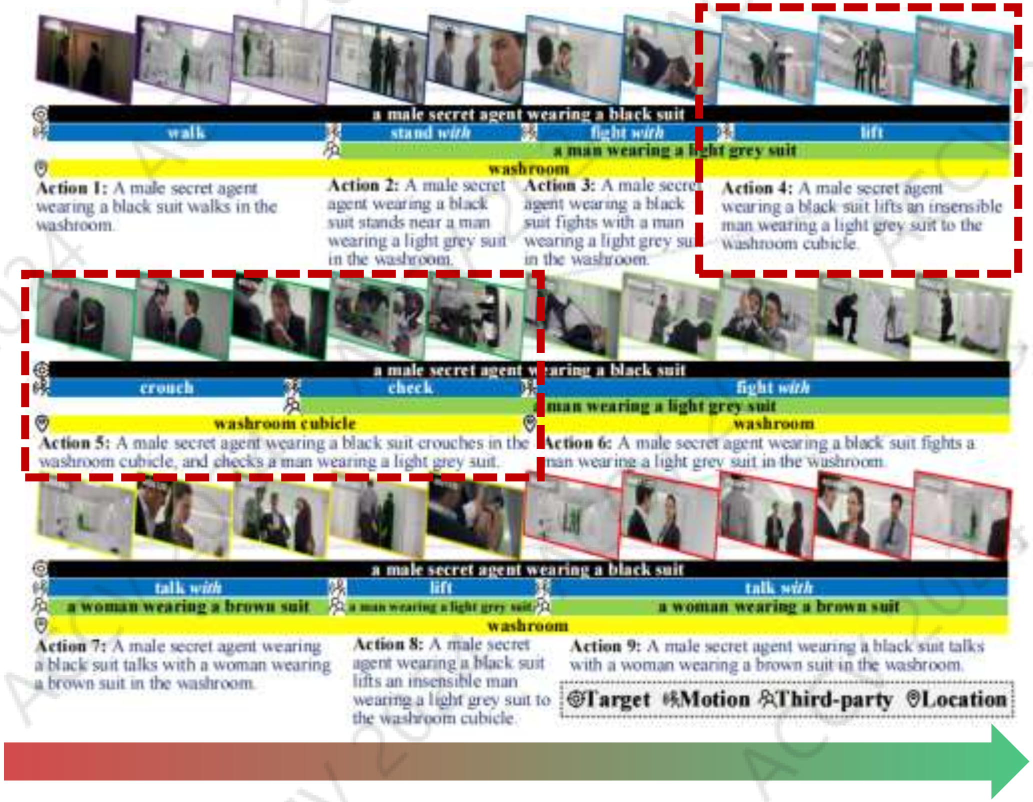
ACCV 2024

Trend 1. More Human-like Task Design

Example: Multimodal Global Instance Tracking (MGIT)

➢ Applying **hierarchical structure inspired by human cognition** for multi-granular annotation

- **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
- **Activity :** Using **causality** as a basis for classification


a male secret agent wearing a black suit
walk
stand *with*
fight *with*
lift
a man wearing a light grey suit
washroom
a male secret agent wearing a black suit
crouch
check
fight *with*
man wearing a light grey suit
washroom cubicle
washroom
a male secret agent wearing a black suit
talk *with*
lift
talk *with*
a woman wearing a brown suit
a man wearing a light grey suit
a woman wearing a brown suit
washroom
Target
Motion
Third-party
Location

Action 1: A male secret agent wearing a black suit walks in the washroom.

Action 2: A male secret agent wearing a black suit stands near a man wearing a light grey suit in the washroom.

Action 3: A male secret agent wearing a black suit fights with a man wearing a light grey suit in the washroom.

Action 4: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 5: A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit.

Action 6: A male secret agent wearing a black suit fights a man wearing a light grey suit in the washroom.

Action 7: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

Action 8: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 9: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

***Action 4：** more suitable as the **Result** of activity 1*

***Action 5：** more suitable as the **Cause** for activity 2*

Cause

Result

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

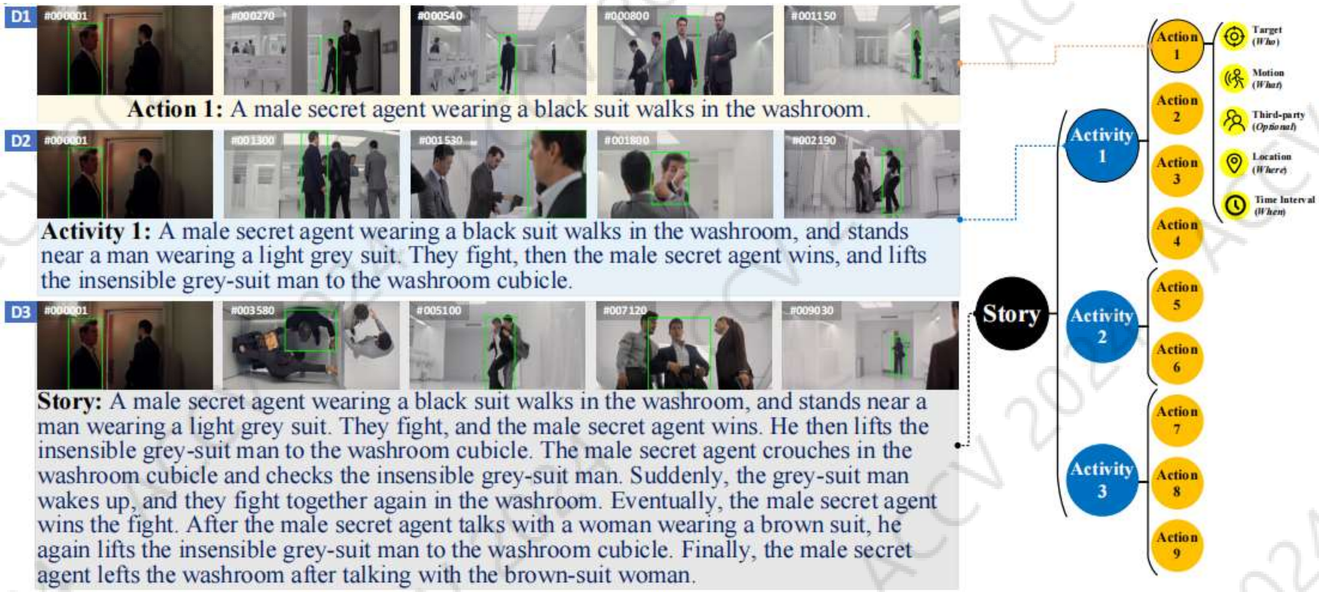
ACCV 2024

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**

➢ Applying **hierarchical structure inspired by human cognition** for multi-granular annotation

  - **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
  - **Activity :** Using **causality** as a basis for classification
  - **Story :** To enhance **temporal and causal relationships**, guiding words such as "first, then, after that, finally," can be used on the basis of actions and activities


D1
#000001
#000270
#000540
#000800
#001150
D2
#000001
#001300
#001530
#001800
#002190
D3
#000001
#003580
#005100
#007120
#009030
Story
Activity 1
Activity 2
Activity 3
Action 1
Action 2
Action 3
Action 4
Action 5
Action 6
Action 7
Action 8
Action 9
Target (Who)
Motion (What)
Third-party (Optional)
Location (Where)
Time Interval (When)

Action 1: A male secret agent wearing a black suit walks in the washroom.

Activity 1: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, then the male secret agent wins, and lifts the insensible grey-suit man to the washroom cubicle.

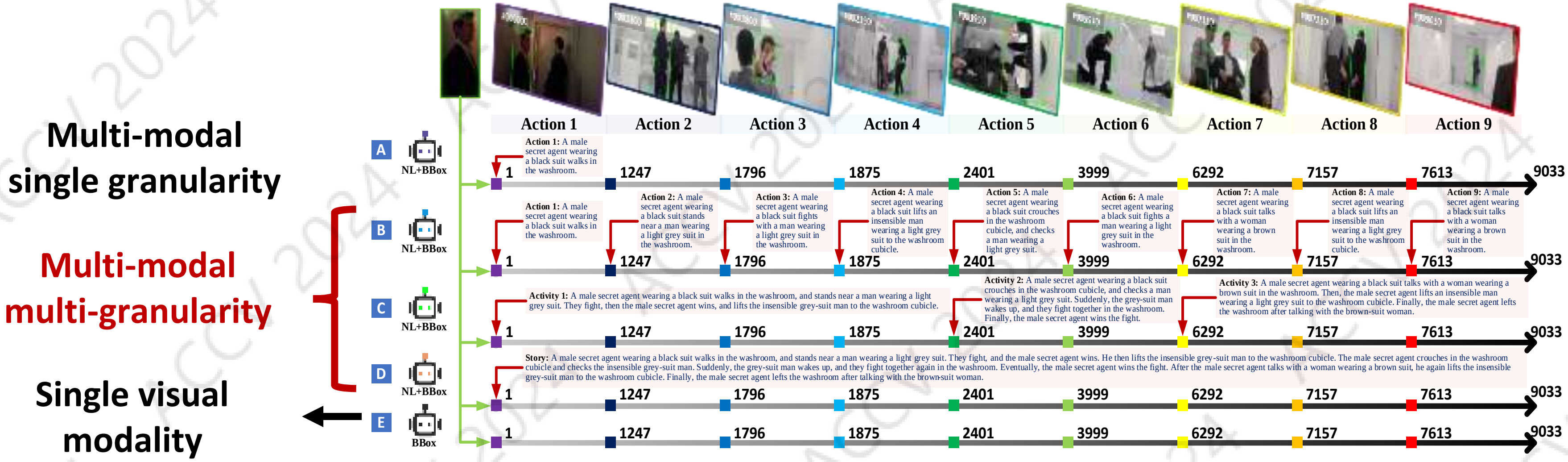
Story: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, and the male secret agent wins. He then lifts the insensible grey-suit man to the washroom cubicle. The male secret agent crouches in the washroom cubicle and checks the insensible grey-suit man. Suddenly, the grey-suit man wakes up, and they fight together again in the washroom. Eventually, the male secret agent wins the fight. After the male secret agent talks with a woman wearing a brown suit, he again lifts the insensible grey-suit man to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**

➢ Expand the evaluation mechanism by conducting experiments under both **traditional evaluation mechanisms** (multi-modal single granularity, single visual modality) and **evaluation mechanisms adapted to this work** (multi-modal multi-granularity).



Multi-modal single granularity

Multi-modal multi-granularity

Single visual modality

Action 1: A male secret agent wearing a black suit walks in the washroom.

Action 1: A male secret agent wearing a black suit walks in the washroom.

Action 2: A male secret agent wearing a black suit stands near a man wearing a light grey suit in the washroom.

Action 3: A male secret agent wearing a black suit fights with a man wearing a light grey suit in the washroom.

Action 4: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 5: A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit.

Action 6: A male secret agent wearing a black suit fights a man wearing a light grey suit in the washroom.

Action 7: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

Action 8: A male secret agent wearing a black suit lifts an insensible man wearing a light grey suit to the washroom cubicle.

Action 9: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom.

Activity 1: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, then the male secret agent wins, and lifts the insensible grey-suit man to the washroom cubicle.

Activity 2: A male secret agent wearing a black suit crouches in the washroom cubicle, and checks a man wearing a light grey suit. Suddenly, the grey-suit man wakes up, and they fight together in the washroom. Finally, the male secret agent wins the fight.

Activity 3: A male secret agent wearing a black suit talks with a woman wearing a brown suit in the washroom. Then, the male secret agent lifts an insensible man wearing a light grey suit to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

Story: A male secret agent wearing a black suit walks in the washroom, and stands near a man wearing a light grey suit. They fight, and the male secret agent wins. He then lifts the insensible grey-suit man to the washroom cubicle. The male secret agent crouches in the washroom cubicle and checks the insensible grey-suit man. Suddenly, the grey-suit man wakes up, and they fight together again in the washroom. Eventually, the male secret agent wins the fight. After the male secret agent talks with a woman wearing a brown suit, he again lifts the insensible grey-suit man to the washroom cubicle. Finally, the male secret agent lefts the washroom after talking with the brown-suit woman.

Adapted evaluation process for different task settings

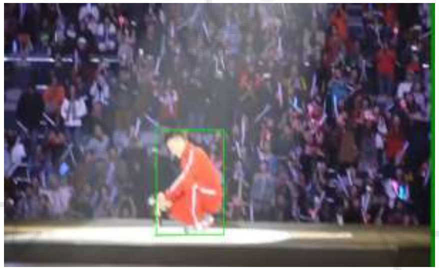
ACCV 2024

*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 1. More Human-like Task Design

- **Example: Multimodal Global Instance Tracking (MGIT)**

➢ Incorporate semantic information into the GIT task and introduced the Multi-modal GIT (MGIT) task → **Visual reasoning** in complex spatio-temporal causal relationships.



A long-term tracking demo

Short-term tracking & Long-term tracking

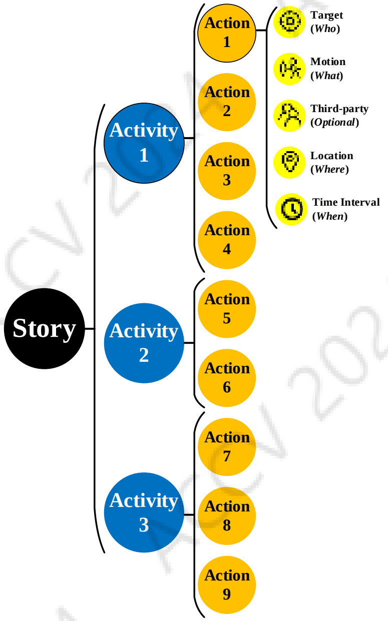
➢ Methods utilize local search to locate the target near to its position in the previous frame → **perceptual level**

No.000676
Shotcut Start: False
Occlusion: True

Global Instance Tracking (GIT)

➢ Methods should remember the target and re-detect it in a new shot→ **cognitive level**

Visual Information


Story
Activity 1
Activity 2
Activity 3
Action 1
Action 2
Action 3
Action 4
Action 5
Action 6
Action 7
Action 8
Action 9
Target (Who)
Motion (What)
Third-party (Optional)
Location (Where)
Time Interval (When)

Semantic Information

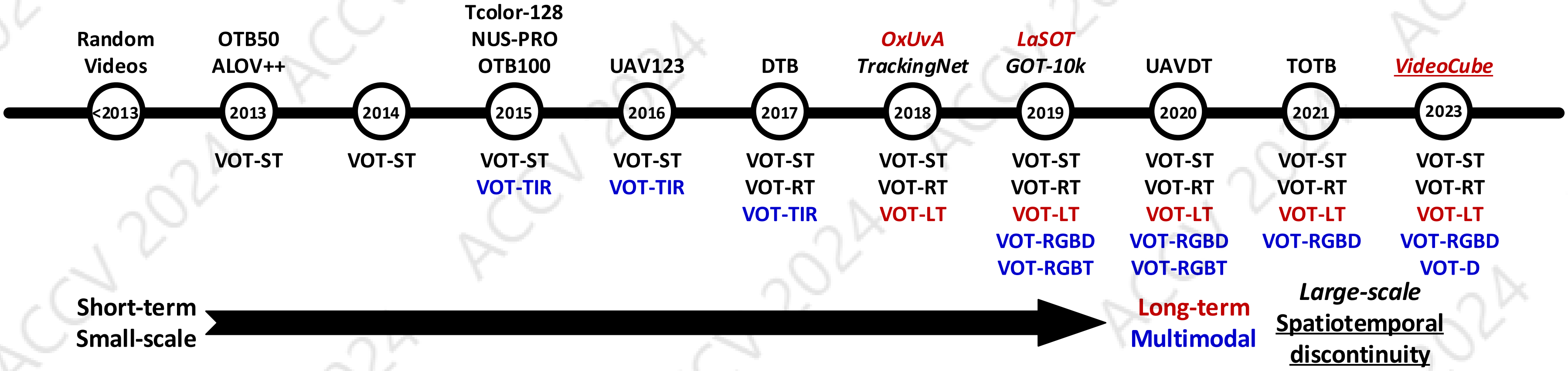
*S. Hu, D. Zhang, M. Wu, et al., "A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship," in the 37th Conference on Neural Information Processing Systems (**NeurIPS**), 2023.*

Trend 2. More Realistic Data Environment

ACCV 2024

Trend 2. More Realistic Data Environment

- What are the living **environments** of humans? → Constructing more comprehensive and realistic datasets.



<2013	Random Videos	
2013	OTB50 ALOV++	VOT-ST
2014		VOT-ST
2015	Tcolor-128 NUS-PRO OTB100	VOT-ST VOT-TIR
2016	UAV123	VOT-ST VOT-TIR
2017	DTB	VOT-ST VOT-RT VOT-TIR
2018	*OxUvA* *TrackingNet*	VOT-ST VOT-RT VOT-LT
2019	*LaSOT* *GOT-10k*	VOT-ST VOT-RT VOT-LT VOT-RGBD VOT-RGBT
2020	UAVDT	VOT-ST VOT-RT VOT-LT VOT-RGBD VOT-RGBT
2021	TOTB	VOT-ST VOT-RT VOT-LT VOT-RGBD
2023	*VideoCube*	VOT-ST VOT-RT VOT-LT VOT-RGBD VOT-D

Short-term
Small-scale
→
Long-term
Multimodal

Large-scale
Spatiotemporal
discontinuity

- ➢ As computer vision technology advances, the need for evaluation environments that closely resemble the real world becomes more important. Realism in evaluation environments refers to **simulating the dynamic, diverse, and unpredictable nature of the real world** to assess the algorithm's performance in practical scenarios.

ACCV 2024

Trend 2. More Realistic Data Environment

- **From model-centric to data-centric**

nature machine intelligence

PERSPECTIVE

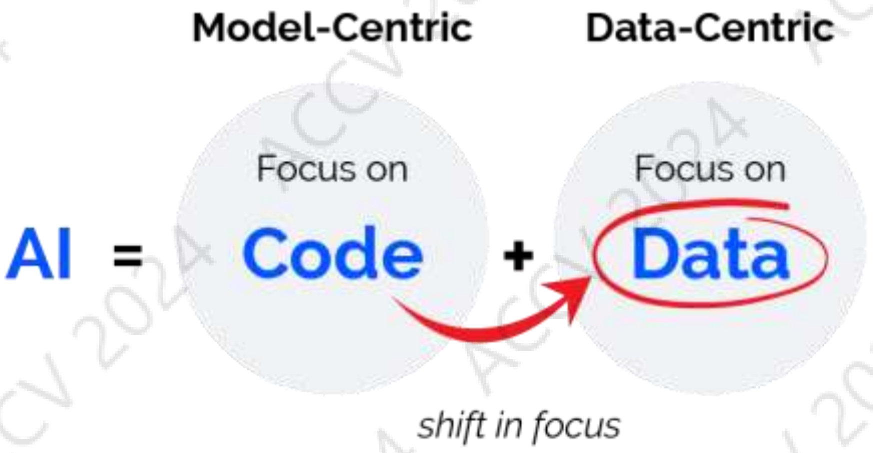
https://doi.org/10.1038/s42256-022-00516-1

Check for updates

Advances, challenges and opportunities in creating data for trustworthy AI

Weixin Liang[1], Girmaw Abebe Tadesse[2], Daniel Ho[3], Fei-Fei Li[1], Matei Zaharia[1], Ce Zhang[4] and James Zou[1,5] ✉

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]IBM Research – Africa, Nairobi, Kenya. [3]Stanford Law School, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ✉e-mail: jamesz@stanford.edu



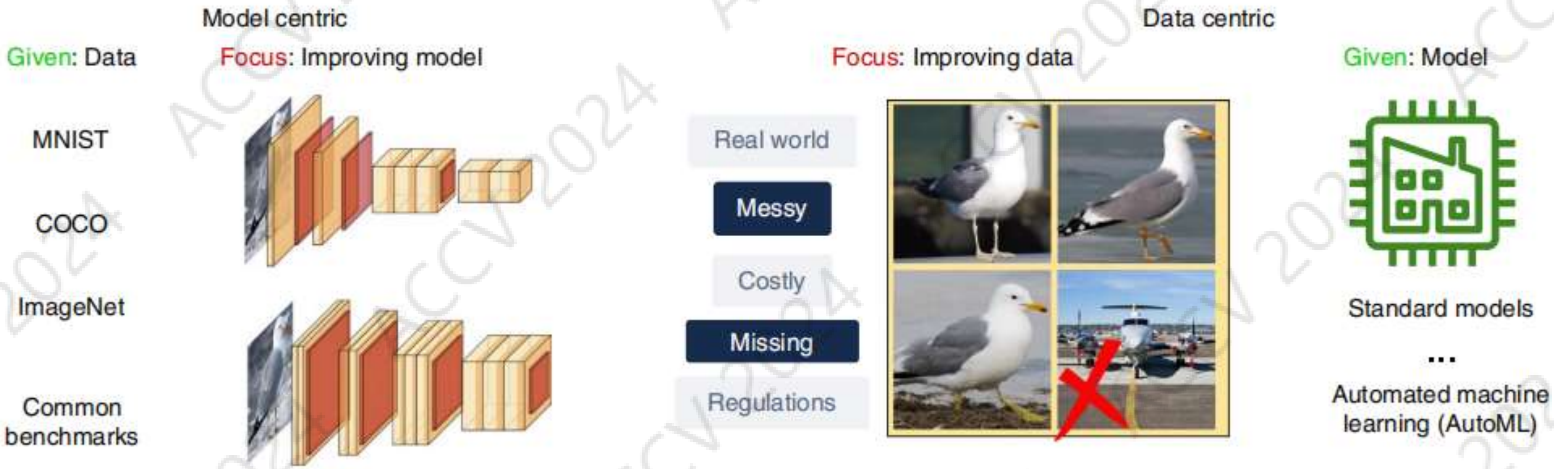
- ➢ More attention needs to be placed on developing methods and standards to improve the **data-for-AI pipeline**.

Liang W, Tadesse G A, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. Nature Machine Intelligence, 2022, 4(8): 669-677.

ACCV 2024

Trend 2. More Realistic Data Environment

- **From model-centric to data-centric**


Model centric
Given: Data
Focus: Improving model
MNIST
COCO
ImageNet
Common benchmarks
Data centric
Focus: Improving data
Given: Model
Real world
Messy
Costly
Missing
Regulations
Standard models
...
Automated machine learning (AutoML)

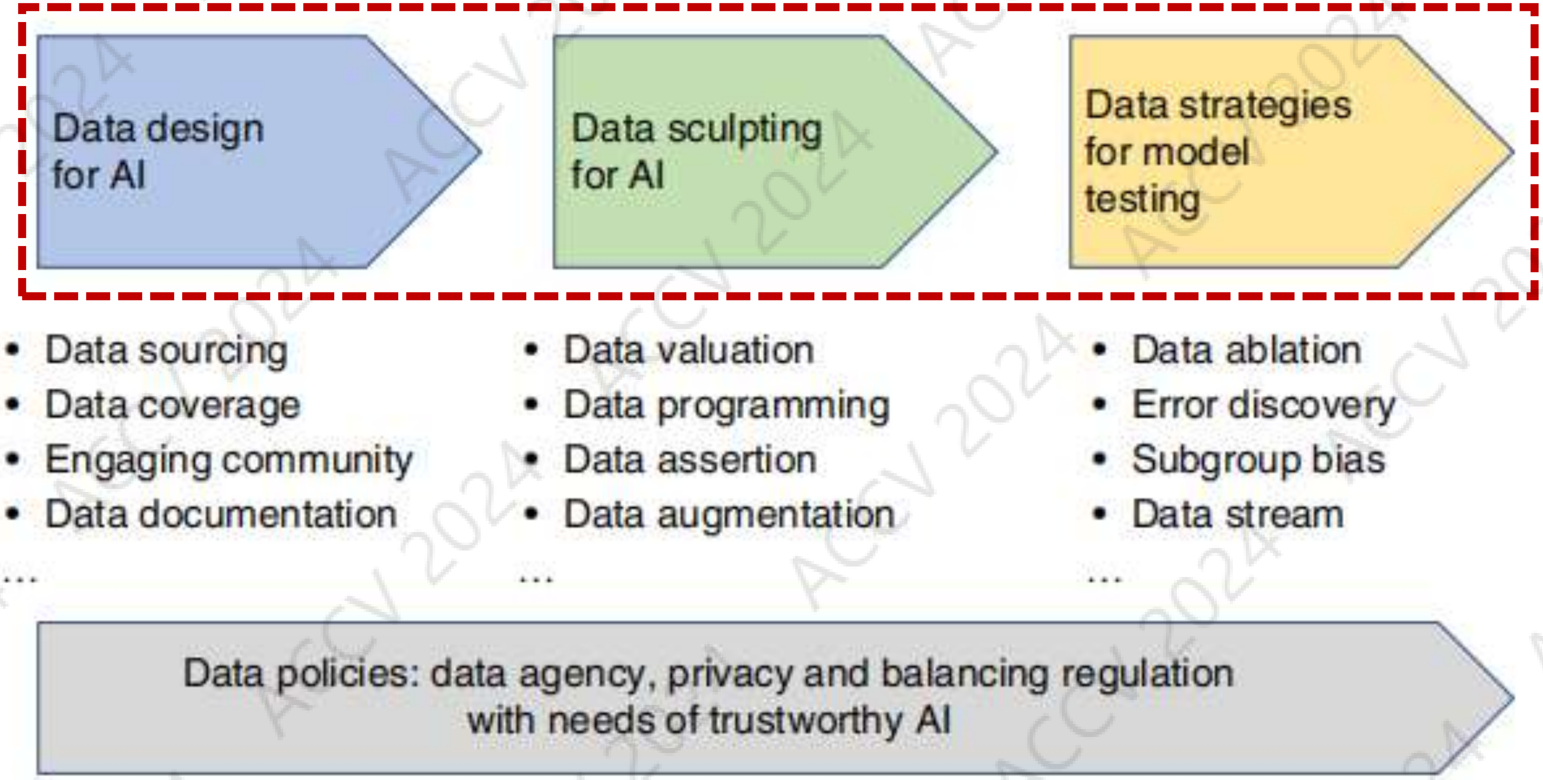
- ➢ **Model-centric** research typically considers data as given and focuses on improving the model architecture or optimization on this data.
- ➢ **Data-centric** research focuses on scalable methods to systematically **improve the data pipeline with data cleaning, selection, annotation and so on**.

Liang W, Tadesse G A, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. Nature Machine Intelligence, 2022, 4(8): 669-677.

ACCV 2024

Trend 2. More Realistic Data Environment

- **From model-centric to data-centric**
  - **Key Steps:**
    - Step 1. Dataset **design** for AI
    - Step 2. Data **sculpting** for AI
    - Step 3. Data strategies for **model testing**



Data design for AI

Data sculpting for AI

Data strategies for model testing

- Data sourcing
- Data coverage
- Engaging community
- Data documentation

...

- Data valuation
- Data programming
- Data assertion
- Data augmentation

...

- Data ablation
- Error discovery
- Subgroup bias
- Data stream

...

Data policies: data agency, privacy and balancing regulation with needs of trustworthy AI

ACCV 2024

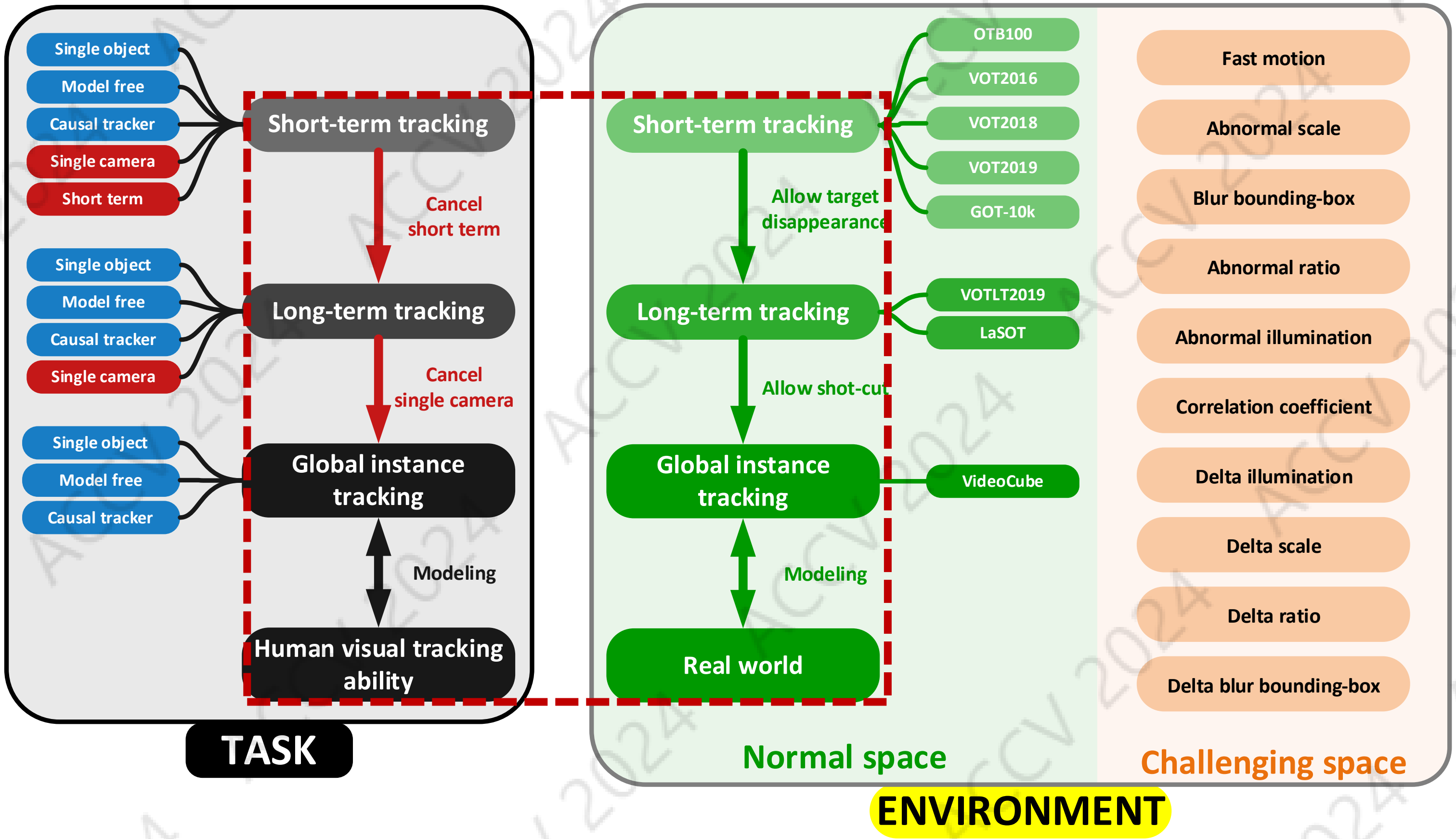
Liang W, Tadesse G A, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. Nature Machine Intelligence, 2022, 4(8): 669-677.

Trend 2. More Realistic Data Environment

ACCV

Step 1. Data design for AI

- Once an **AI application** has been identified, **designing the data**—namely identifying and documenting the sources of data—to develop the AI model is often one of the first considerations.


Single object
Model free
Causal tracker
Single camera
Short term
Short-term tracking
Cancel short term
Single object
Model free
Causal tracker
Single camera
Long-term tracking
Cancel single camera
Single object
Model free
Causal tracker
Global instance tracking
Modeling
Human visual tracking ability
TASK
Short-term tracking
OTB100
VOT2016
VOT2018
VOT2019
GOT-10k
Allow target disappearance
Long-term tracking
VOTLT2019
LaSOT
Allow shot-cut
Global instance tracking
VideoCube
Modeling
Real world
Normal space
Fast motion
Abnormal scale
Blur bounding-box
Abnormal ratio
Abnormal illumination
Correlation coefficient
Delta illumination
Delta scale
Delta ratio
Delta blur bounding-box
Challenging space
ENVIRONMENT

Trend 2. More Realistic Data Environment

Step 1. Data design for AI

- Design should be an **iterative** process—it is often useful to have pilot data to develop an initial AI model and then **collect additional data** to patch the model's limitations.

GOT-10k

2017.11 ● **CVPR Submit**
1,000 videos

2018.10 ● **TPAMI Submit**
10,000 videos
Open-set evaluation
Generalization

VideoCube

2020.03 ● **ECCV Submit**
GIT task
250 videos

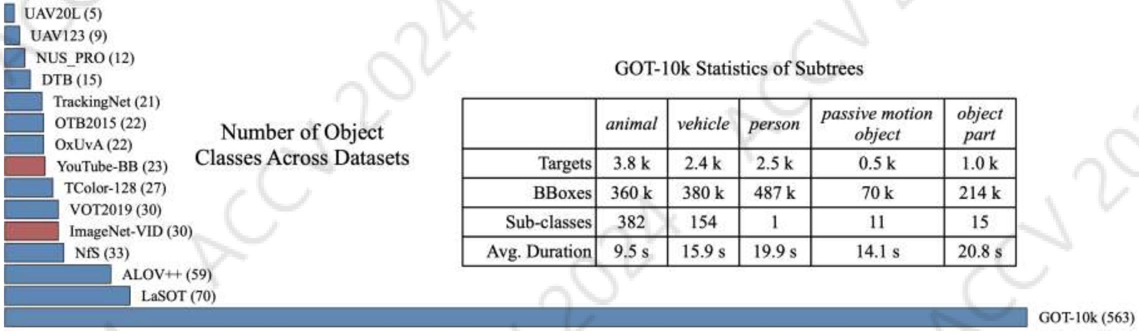
2021.05 ● **TPAMI Submit**
GIT task
500 videos
Eye-tracking

Tiny version → Official version

ACCV 2024

Trend 2. More Realistic Data Environment

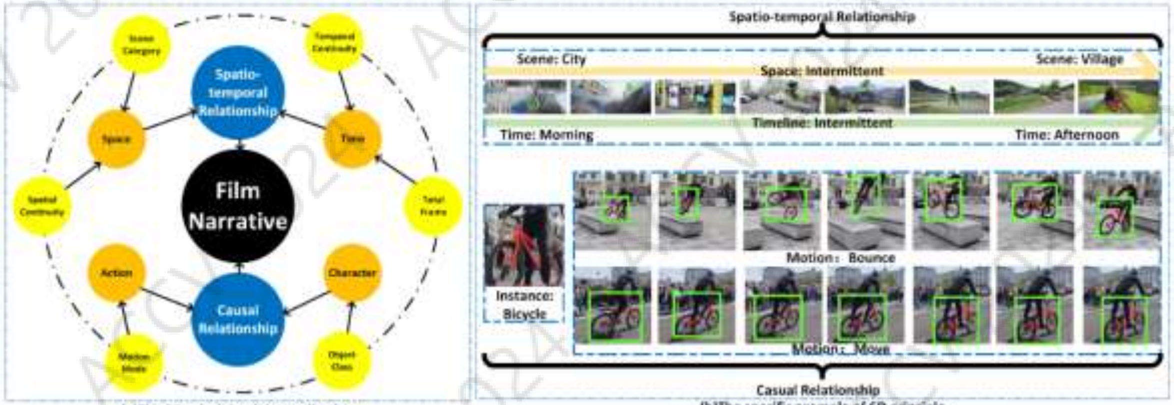
- **Step 1. Data design for AI**
  - A critical design criterion is to ensure that the data are appropriate for the task and have good **coverage to represent diverse users and scenarios** that the model can encounter in practice.



GOT-10k Statistics of Subtrees

	animal	*vehicle*	*person*	*passive motion object*	*object part*
Targets	3.8 k	2.4 k	2.5 k	0.5 k	1.0 k
BBoxes	360 k	380 k	487 k	70 k	214 k
Sub-classes	382	154	1	11	15
Avg. Duration	9.5 s	15.9 s	19.9 s	14.1 s	20.8 s

GOT-10K: 563 object classes， based on WordNet


Scene Category
Temporal Continuity
Spatio-temporal Relationship
Space
Time
Film Narrative
Spatial Continuity
Total Frame
Action
Character
Causal Relationship
Motion Mode
Object Class
Spatio-temporal Relationship
Scene: City
Scene: Village
Space: Intermittent
Timeline: Intermittent
Time: Morning
Time: Afternoon
Instance: Bicycle
Motion： Bounce
Motion： Move
Casual Relationship

(a)The 6D principle of data collection

(b)The specific example of 6D principle

VideoCube: use 6D principle to model the real scenarios

Trend 2. More Realistic Data Environment

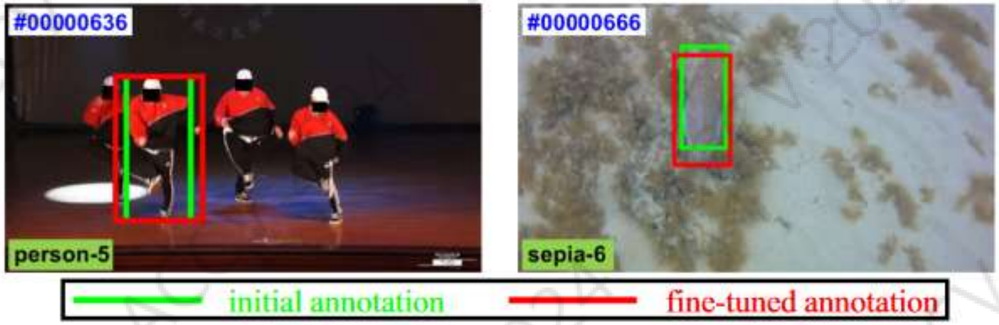
- **Step 1. Data design for AI**
  - When representative data are hard to access, **synthetic data** can potentially fill some of the coverage gaps.

UAV123: Rotary-wing UAV (DJI S1000) + UAV simulator (UE4)

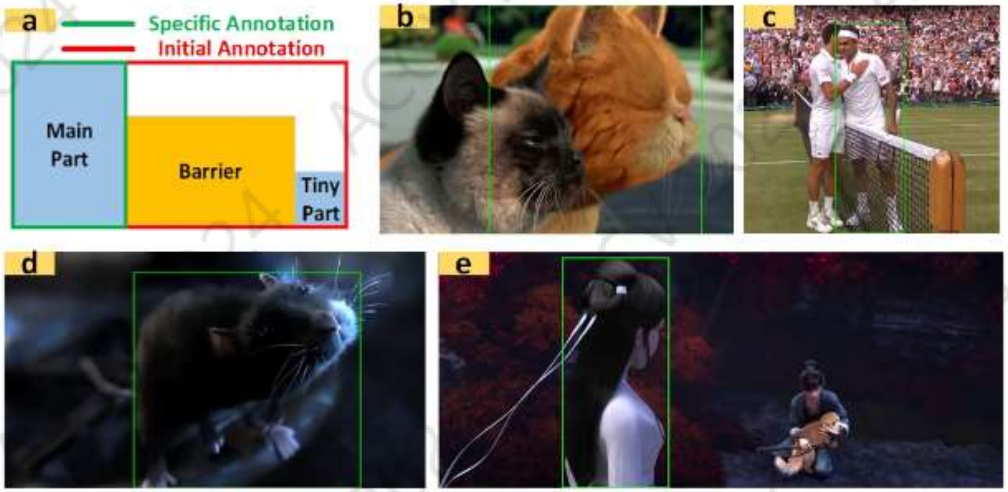
ACCV 2024

Trend 2. More Realistic Data Environment

- **Step 2. Data sculpting for AI**
  - Once an initial dataset is collected, a substantial amount of work is needed to **sculpt or refine** the data to make it **effective** for AI development.


#00000636
person-5
#00000666
sepia-6
initial annotation
fine-tuned annotation

LaSOT: fine-tuning initial annotations


a
Specific Annotation
Initial Annotation
Main Part
Barrier
Tiny Part
b
c
d
e

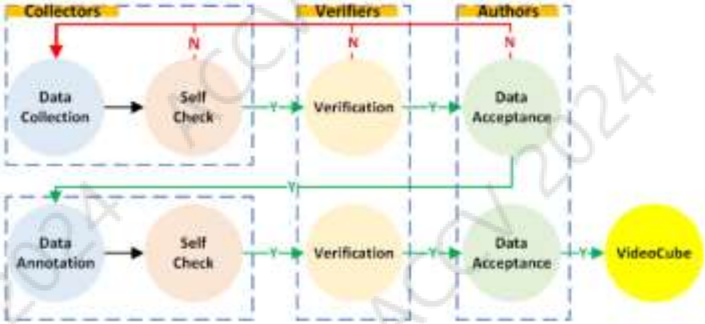
VideoCube: specific annotation rules

ACCV 2024

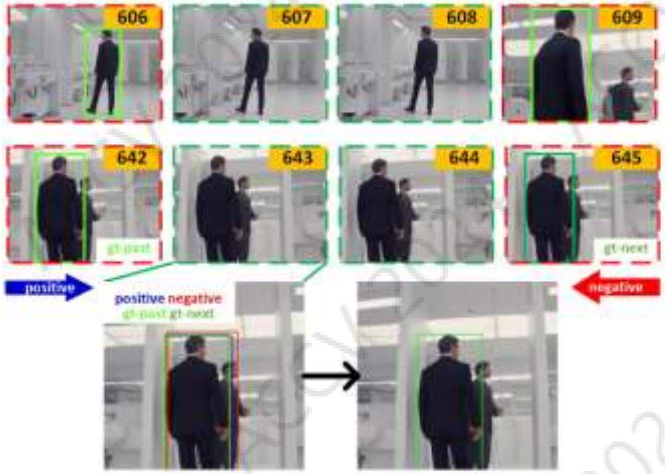
Trend 2. More Realistic Data Environment

Step 2. Data sculpting for AI

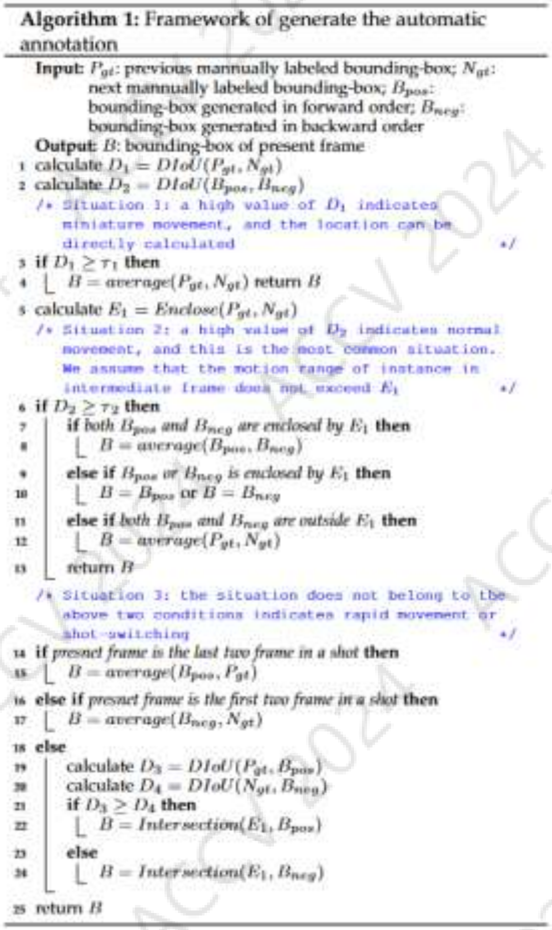
- A human-in-the-loop approach to reduce annotation costs is to prioritize the **most valuable data for humans to annotate**.


Collectors
Verifiers
Authors
N
N
N
Data Collection
Self Check
Y
Verification
Y
Data Acceptance
Y
Data Annotation
Self Check
Y
Verification
Y
Data Acceptance
Y
VideoCube

VideoCube: data verification process


606
607
608
609
642
643
644
645
gt-past
gt-next
positive
negative
positive negative
gt-past gt-next

VideoCube: automatic annotation



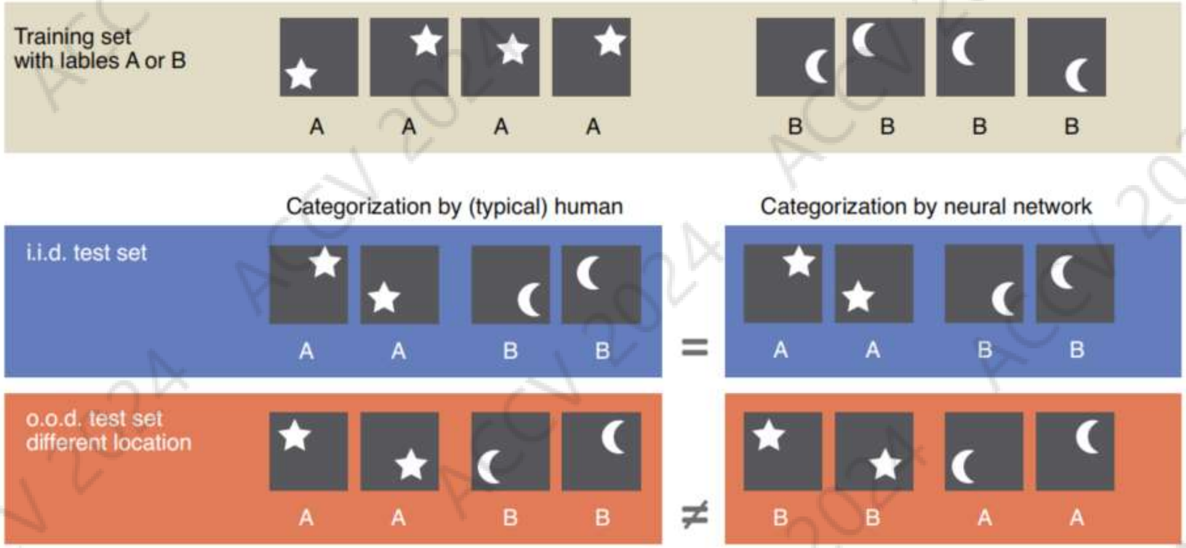
```
Algorithm 1: Framework of generate the automatic annotation
Input: $P_{gt}$: previous mannually labeled bounding-box; $N_{gt}$: next mannually labeled bounding-box; $B_{pos}$: bounding-box generated in forward order; $B_{neg}$: bounding-box generated in backward order
Output: $B$: bounding-box of present frame
calculate $D_1 = DIoU(P_{gt}, N_{gt})$
calculate $D_2 = DIoU(B_{pos}, B_{neg})$
/* Situation 1: a high value of $D_1$ indicates miniature movement, and the location can be directly calculated */
if $D_1 \geq \tau_1$ then
    $B = average(P_{gt}, N_{gt})$ return $B$
calculate $E_1 = Enclose(P_{gt}, N_{gt})$
/* Situation 2: a high value of $D_2$ indicates normal movement, and this is the most common situation. We assume that the motion range of instance in intermediate frame does not exceed $E_1$ */
if $D_2 \geq \tau_2$ then
    if both $B_{pos}$ and $B_{neg}$ are enclosed by $E_1$ then
        $B = average(B_{pos}, B_{neg})$
    else if $B_{pos}$ or $B_{neg}$ is enclosed by $E_1$ then
        $B = B_{pos}$ or $B = B_{neg}$
    else if both $B_{pos}$ and $B_{neg}$ are outside $E_1$ then
        $B = average(P_{gt}, N_{gt})$
    return $B$
/* Situation 3: the situation does not belong to the above two conditions indicates rapid movement or shot-switching */
if presnet frame is the last two frame in a shot then
    $B = average(B_{pos}, P_{gt})$
else if presnet frame is the first two frame in a shot then
    $B = average(B_{neg}, N_{gt})$
else
    calculate $D_3 = DIoU(P_{gt}, B_{pos})$
    calculate $D_4 = DIoU(N_{gt}, B_{neg})$
    if $D_3 \geq D_4$ then
        $B = Intersection(E_1, B_{pos})$
    else
        $B = Intersection(E_1, B_{neg})$
return $B$
```

ACCV 2024

Trend 2. More Realistic Data Environment

Step 3. Data strategies for model testing

- An important aspect of evaluation is to verify that the AI models **do not use ‘shortcut’ strategies**.


Training set
with lables A or B
A
A
A
A
B
B
B
B
Categorization by (typical) human
Categorization by neural network
i.i.d. test set
A
A
B
B
=
A
A
B
B
o.o.d. test set
different location
A
A
B
B
≠
B
B
A
A

An example of shortcut

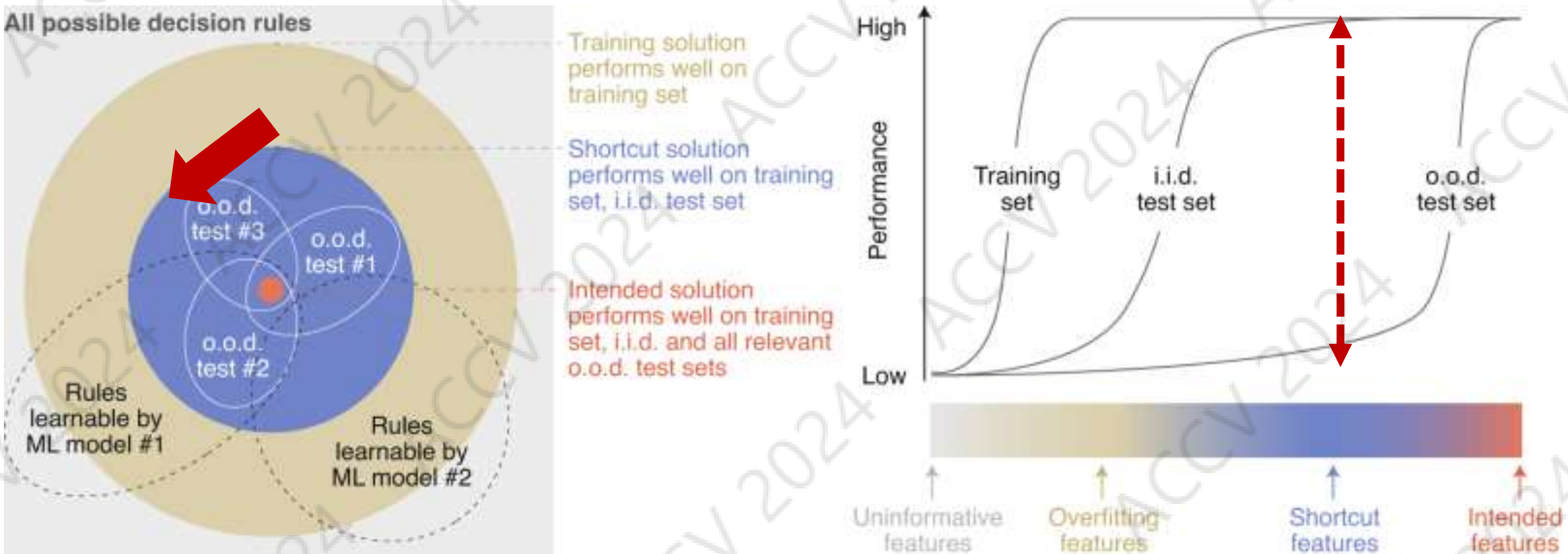
When trained on a simple dataset of stars and moons, a standard fully connected neural network learns a shortcut strategy: **classifying based on the location** (stars in the top right or bottom left; moons in the top left or bottom right) **rather than the shape of the objects.**

Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks[J]. Nature Machine Intelligence, 2020, 2(11): 665-673.

ACCV 2024

Trend 2. More Realistic Data Environment

- ## Step 3. Data strategies for model testing



All possible decision rules
Training solution performs well on training set
Shortcut solution performs well on training set, i.i.d. test set
Intended solution performs well on training set, i.i.d. and all relevant o.o.d. test sets
o.o.d. test #3
o.o.d. test #1
o.o.d. test #2
Rules learnable by ML model #1
Rules learnable by ML model #2
High
Performance
Low
Training set
i.i.d. test set
o.o.d. test set
Uninformative features
Overfitting features
Shortcut features
Intended features

- **I.i.d. test solutions, including shortcuts:** Decision rules that solve both the training and i.i.d. test set typically score high on standard benchmarks, but may fail in o.o.d test set.

- **By counting the number of white pixels (moons are smaller than stars)**
- **By location**
- **By shape**

Shortcuts are decision rules that perform well on i.i.d. test data but fail on o.o.d. tests, revealing a mismatch between intended and learned solution.

Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks[J]. Nature Machine Intelligence, 2020, 2(11): 665-673.

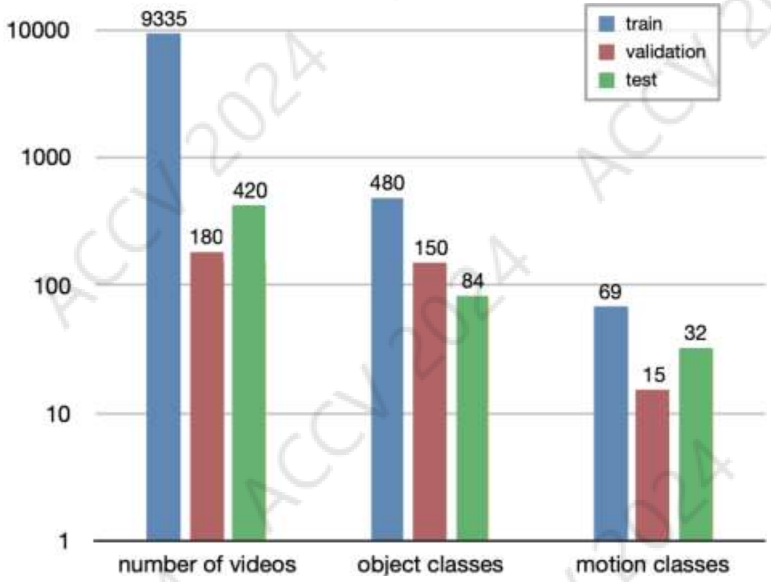
ACCV 2024

Trend 2. More Realistic Data Environment

- **Step 3. Data strategies for model testing**

➢ Towards **o.o.d. generalization** tests for detecting shortcuts:

- If model performance is assessed only on i.i.d. test data, we cannot tell whether the model is actually acquiring the ability we think it is, since **exploiting shortcuts often leads to deceptively good results** on standard metrics.
- **o.o.d. generalization tests should become a standard method for benchmarking models.**



GOT-10k uses open-set evaluation (no overlap between training and testing categories) for generalization ability evaluation

ACCV 2024

Trend 2. More Realistic Data Environment

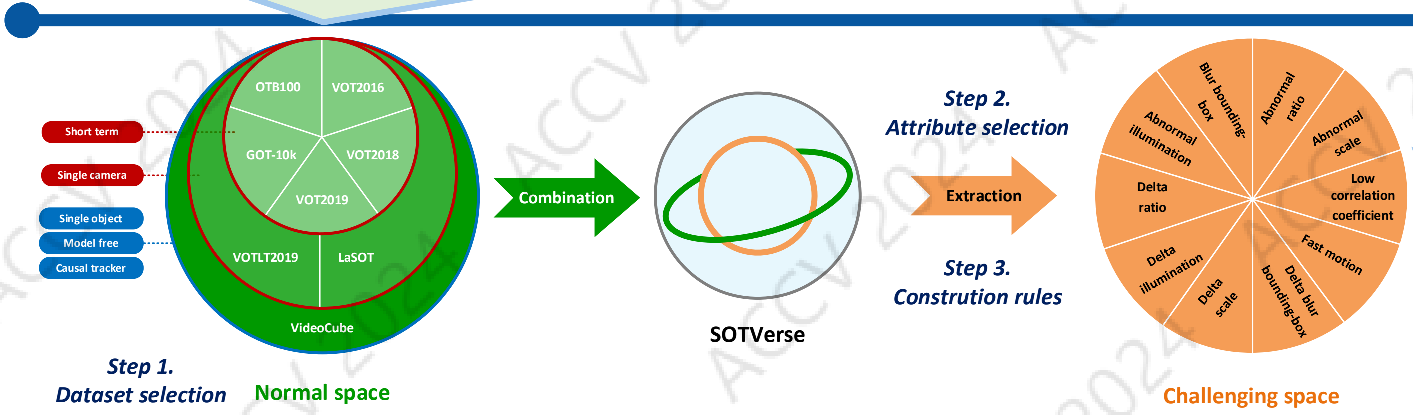
• Example 1: SOTVerse (Dynamic and Open Task Space for VOT)

➢ Motivation

OTB-2013/2015
GOT-10k
VOT2016&2018
VOTLT2019
LaSOT
VideoCube

Existing datasets:

- **Static and closed** after construction
- **Ignore challenging factors**


Short term
Single camera
Single object
Model free
Causal tracker
OTB100
VOT2016
GOT-10k
VOT2018
VOT2019
VOTLT2019
LaSOT
VideoCube
Step 1.
Dataset selection
Normal space
Combination
SOTVerse
Step 2.
Attribute selection
Extraction
Step 3.
Constrution rules
Blur bounding-box
Abnormal ratio
Abnormal scale
Low correlation coefficient
Fast motion
Delta blur bounding-box
Delta scale
Delta illumination
Delta ratio
Abnormal illumination
Challenging space

Low Light

Scale Variation

Motion Blur

➢ Integrate diverse environments to create SOTVerse, a **dynamic and open task space** comprising 12.56 million frames.

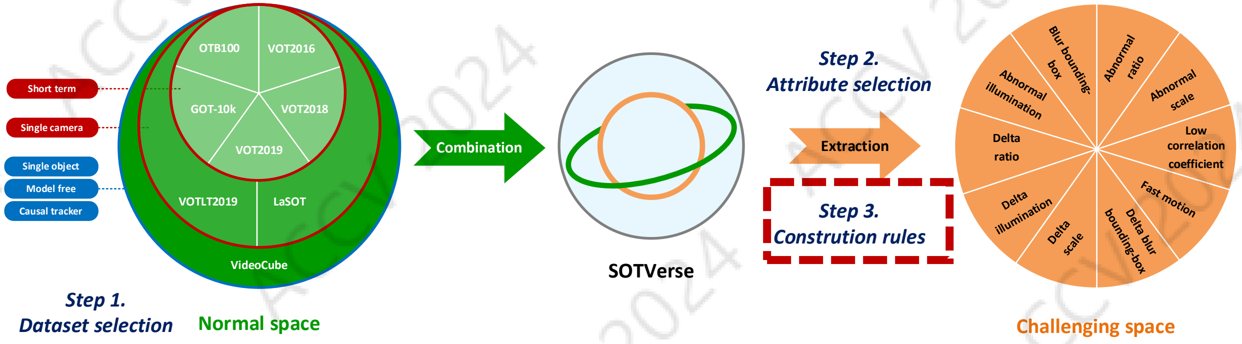
➢ Within this task space, researchers can efficiently construct different subspaces to train algorithms, thereby improving their **visual generalization** across various scenarios.

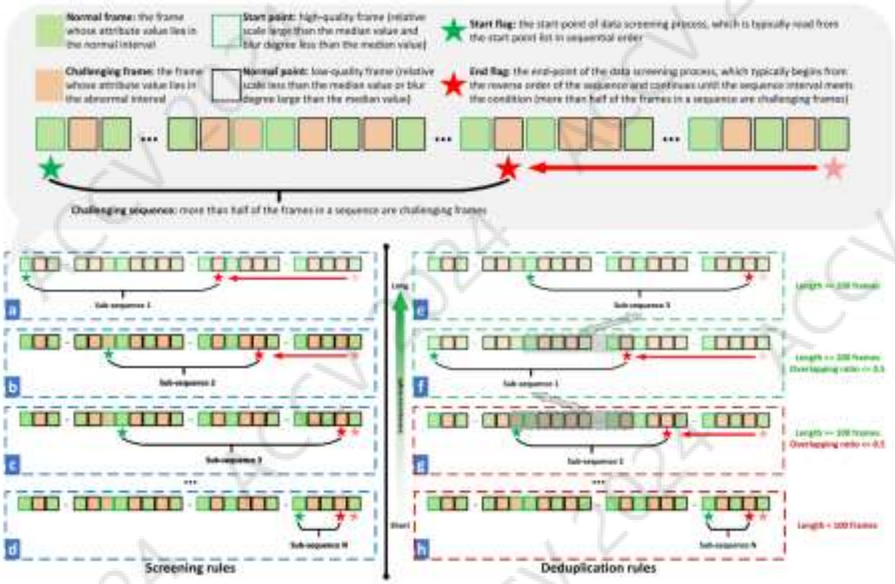
S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (IJCV), 2024.

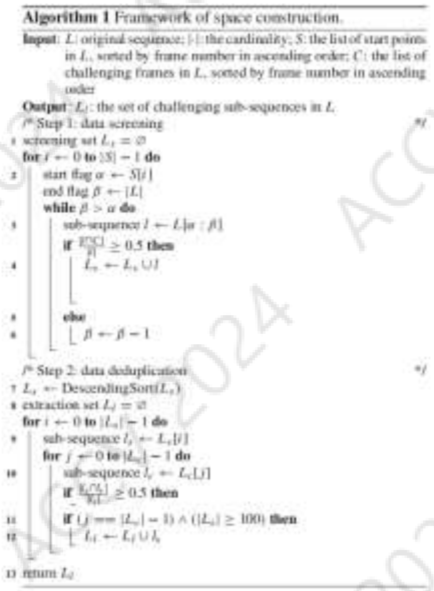
ACCV 2024

Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**
  - ➢ **Automatically mine challenging subsequences** that meet the requirements based on the research goal.


Short term
Single camera
Single object
Model free
Causal tracker
OTB100
VOT2016
GOT-10k
VOT2018
VOT2019
VOTLT2019
LaSOT
VideoCube
Step 1.
Dataset selection
Normal space
Combination
SOTVerse
Step 2.
Attribute selection
Extraction
Step 3.
Constrution rules
Blur bounding-box
Abnormal ratio
Abnormal scale
Low correlation coefficient
Fast motion
Delta blur bounding-box
Delta scale
Delta illumination
Delta ratio
Abnormal illumination
Challenging space


Normal frame: the frame whose attribute value lies in the normal interval
Challenging frame: the frame whose attribute value lies in the abnormal interval
Start point: high-quality frame (relative scale large than the median value and blur degree less than the median value)
Normal point: low-quality frame (relative scale less than the median value or blur degree large than the median value)
Start flag: the start point of data screening process, which is typically read from the start point list in sequential order
End flag: the end-point of the data screening process, which typically begins from the reverse order of the sequence and continues until the sequence interval meets the condition (more than half of the frames in a sequence are challenging frames)
...
Challenging sequence: more than half of the frames in a sequence are challenging frames
a
b
c
d
e
f
g
h
Long
Short
Screening rules
Deduplication rules



Algorithm 1 Framework of space construction.

```
Input: L: original sequence; |·|: the cardinality; S: the list of start points
       in L, sorted by frame number in ascending order; C: the list of
       challenging frames in L, sorted by frame number in ascending
       order
Output: L_f: the set of challenging sub-sequences in L
/* Step 1: data screening                                              */
screening set $L_s = \varnothing$
for $i \leftarrow 0$ to $|S| - 1$ do
    start flag $\alpha \leftarrow S[i]$
    end flag $\beta \leftarrow |L|$
    while $\beta > \alpha$ do
        sub-sequence $l \leftarrow L[\alpha : \beta]$
        if $\frac{|l \cap C|}{|l|} \geq 0.5$ then
            $L_s \leftarrow L_s \cup l$
        else
            $\beta \leftarrow \beta - 1$
/* Step 2: data deduplication                                          */
$L_s \leftarrow$ DescendingSort($L_s$)
extraction set $L_f = \varnothing$
for $i \leftarrow 0$ to $|L_s| - 1$ do
    sub-sequence $l_i \leftarrow L_s[i]$
    for $j \leftarrow 0$ to $|L_s| - 1$ do
        sub-sequence $l_j \leftarrow L_s[j]$
        if $\frac{|l_i \cap l_j|}{|l_i|} \geq 0.5$ then
        if $(j == |L_s| - 1) \wedge (|l_i| \geq 100)$ then
            $l_f \leftarrow L_f \cup l_i$
return $L_f$
```

*S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (**IJCV**), 2024.*

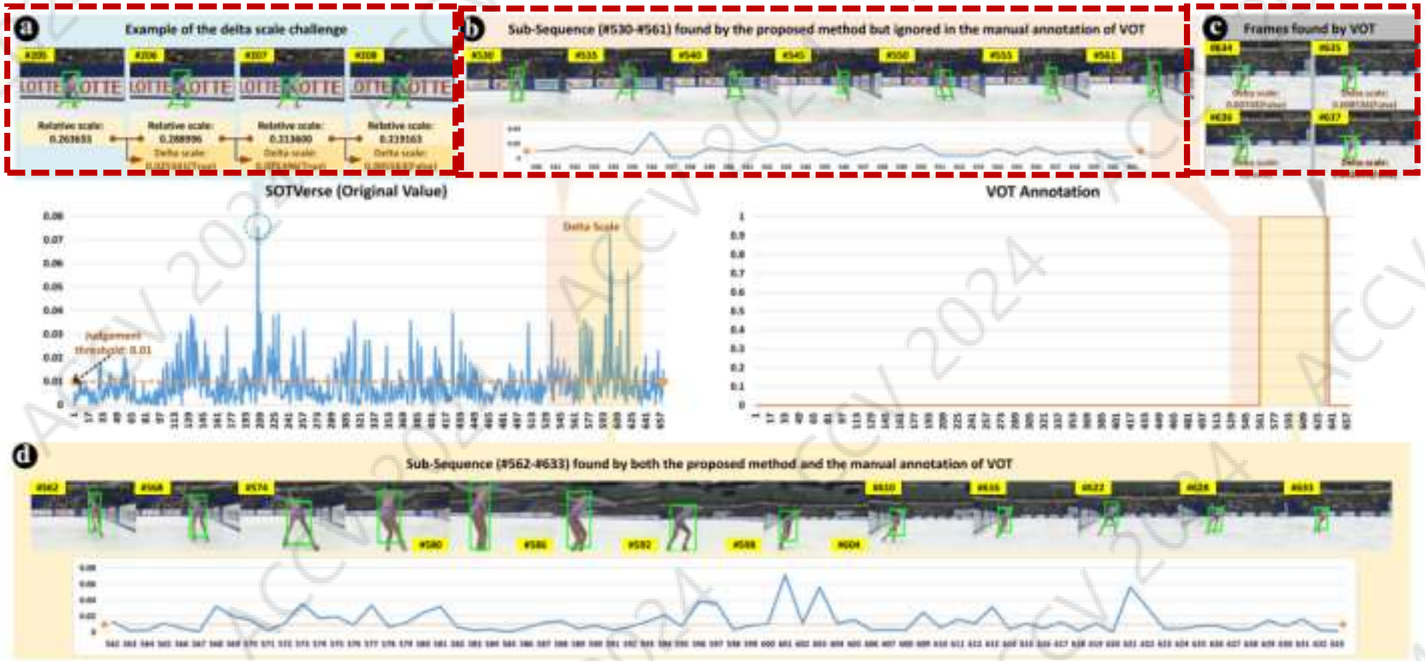
ACCV 2024

Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**

➢ **Comparison with human manual annotations**: Subspace construction strategy can effectively mine highly challenging sequences

- Efficiently focus on **sparsely distributed** challenging video frames
- Effectively mine challenging sequences **ignored by human manual annotation**
- **More accurate judgment** on the starting and ending points of highly challenging sequences

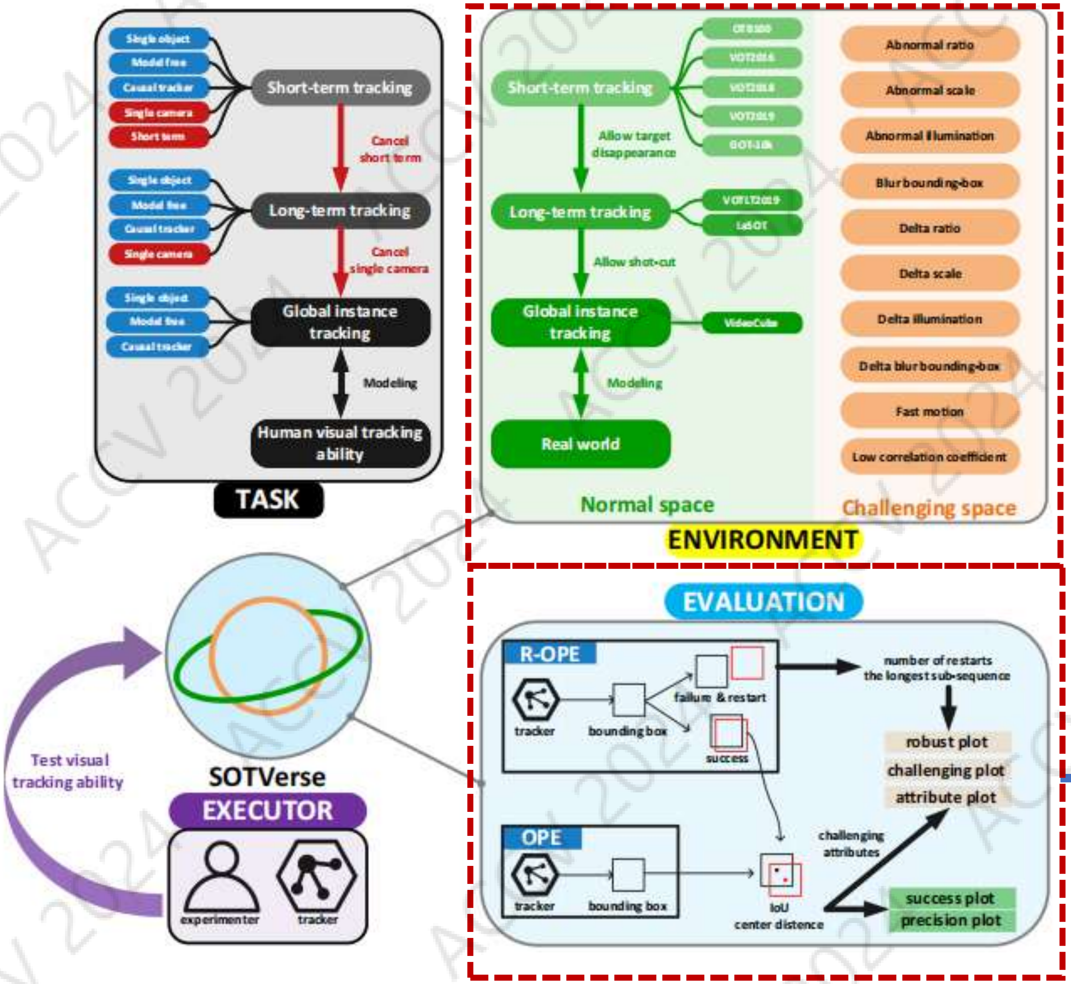


S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (IJCV), 2024.

ACCV 2024

Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**


Single object
Model free
Causal tracker
Single camera
Short term
Short-term tracking
Cancel short term
Single object
Model free
Causal tracker
Single camera
Long-term tracking
Cancel single camera
Single object
Model free
Causal tracker
Global instance tracking
Modeling
Human visual tracking ability
TASK
Short-term tracking
OTB100
VOT2016
VOT2018
VOT2019
GOT-10k
Allow target disappearance
Long-term tracking
VOTLT2019
LaSOT
Allow shot-cut
Global instance tracking
VideoCube
Modeling
Real world
Normal space
Abnormal ratio
Abnormal scale
Abnormal illumination
Blur bounding-box
Delta ratio
Delta scale
Delta illumination
Delta blur bounding-box
Fast motion
Low correlation coefficient
Challenging space
ENVIRONMENT
EVALUATION
R-OPE
tracker
bounding box
failure & restart
success
number of restarts
the longest sub-sequence
robust plot
challenging plot
attribute plot
OPE
tracker
bounding box
IoU
center distence
challenging attributes
success plot
precision plot
Test visual tracking ability
SOTVerse
EXECUTOR
experimenter
tracker

➢ **Some key issues that are masked by traditional evaluation methods:**
  - How does the SOTA algorithm perform when faced with difficult frames?
  - To which difficult challenges are algorithms more susceptible?
  - How well does the algorithm track over long sequences?

We **get nothing** by only focusing on scores under traditional evaluation methods.

*S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (**IJCV**), 2024.*

ACCV 2024

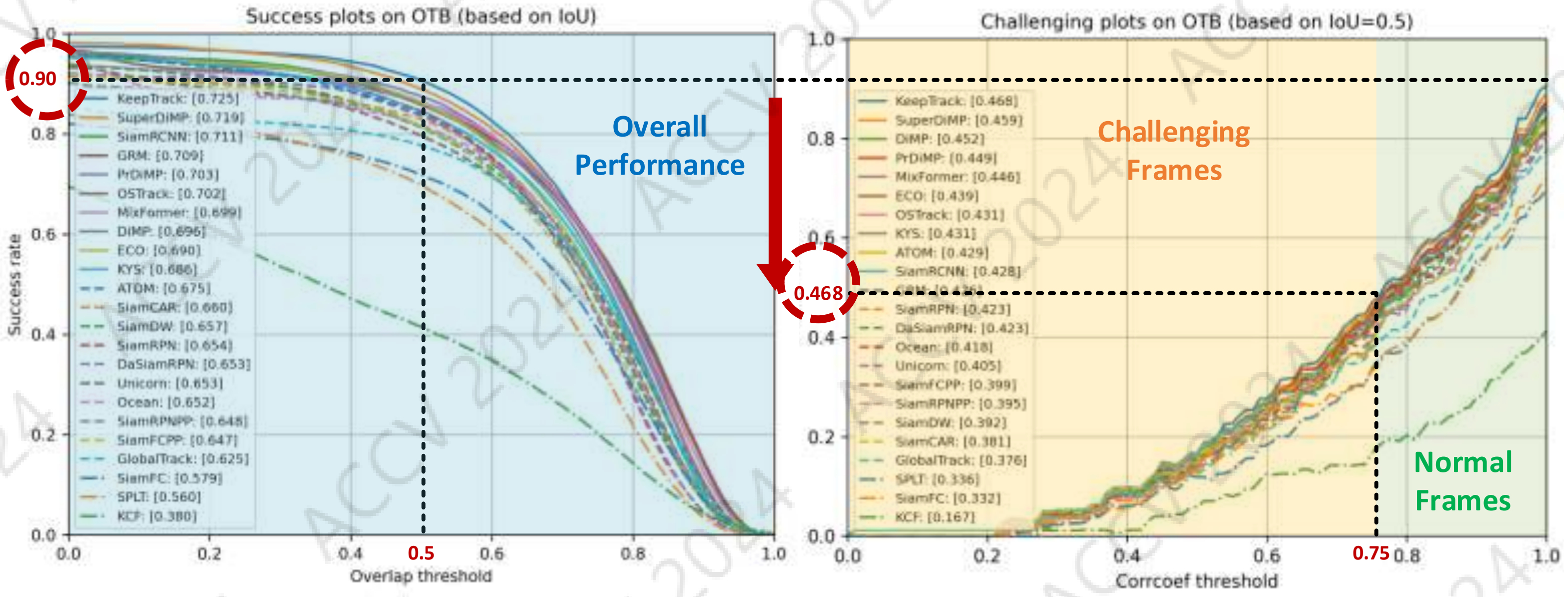
Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**

➢ **Some key issues that are masked by traditional evaluation methods:** How does the SOTA algorithm perform when faced with difficult frames?

  - Difficult frame: correlation coefficient between two frames <0.75
  - **Challenge plot:** Calculates the success rate of the algorithm over all difficult frames

    ☐ **The averaging form used in existing evaluation metrics will mask the bottleneck of the algorithm's ability on difficult frames.**

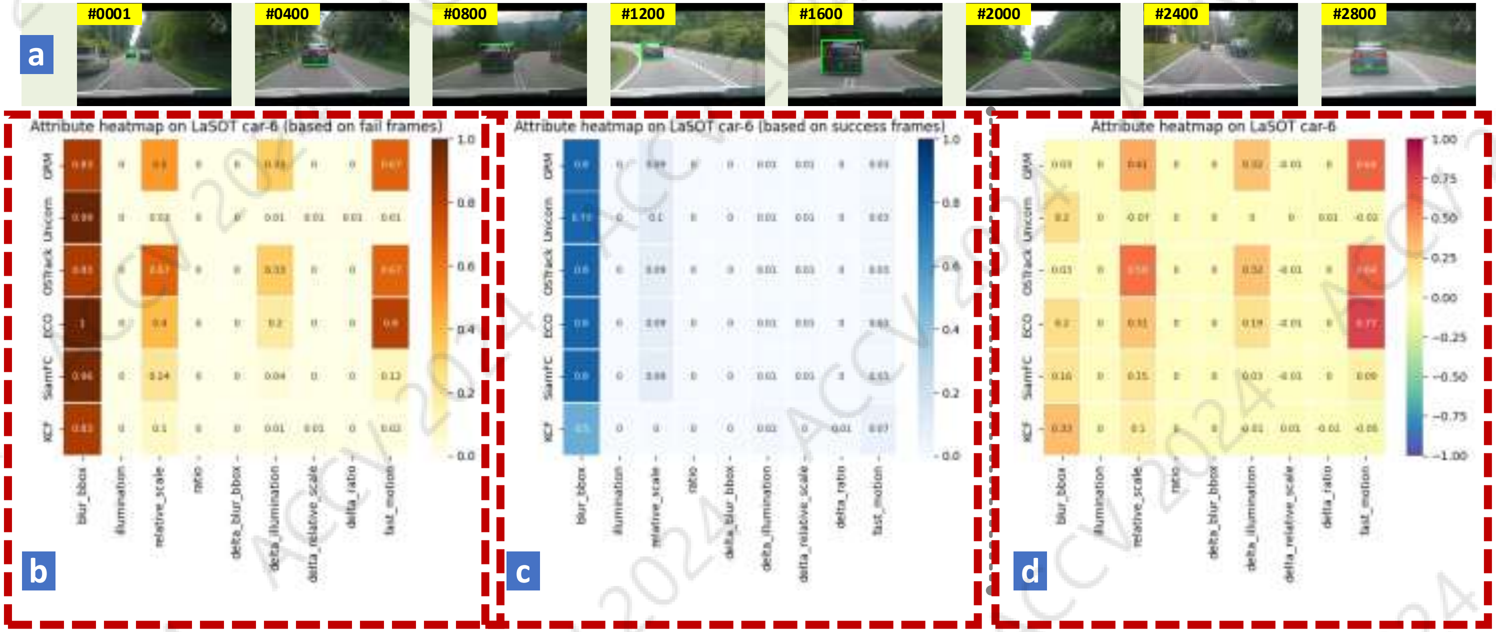


*S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (**IJCV**), 2024.*

Trend 2. More Realistic Data Environment

Example 1: SOTVerse (Dynamic and Open Task Space for VOT)

- **Some key issues that are masked by traditional evaluation methods:** To which difficult challenges are algorithms more susceptible?
  - Failure frame: frame where algorithm tracking fails (IoU<0.5)
  - Successful frame: frame where the algorithm successfully tracks (IoU>=0.5)
  - **Attribute plot:** Find the attribute with the largest difference between the failure frame and the success frame



ACCV 2024

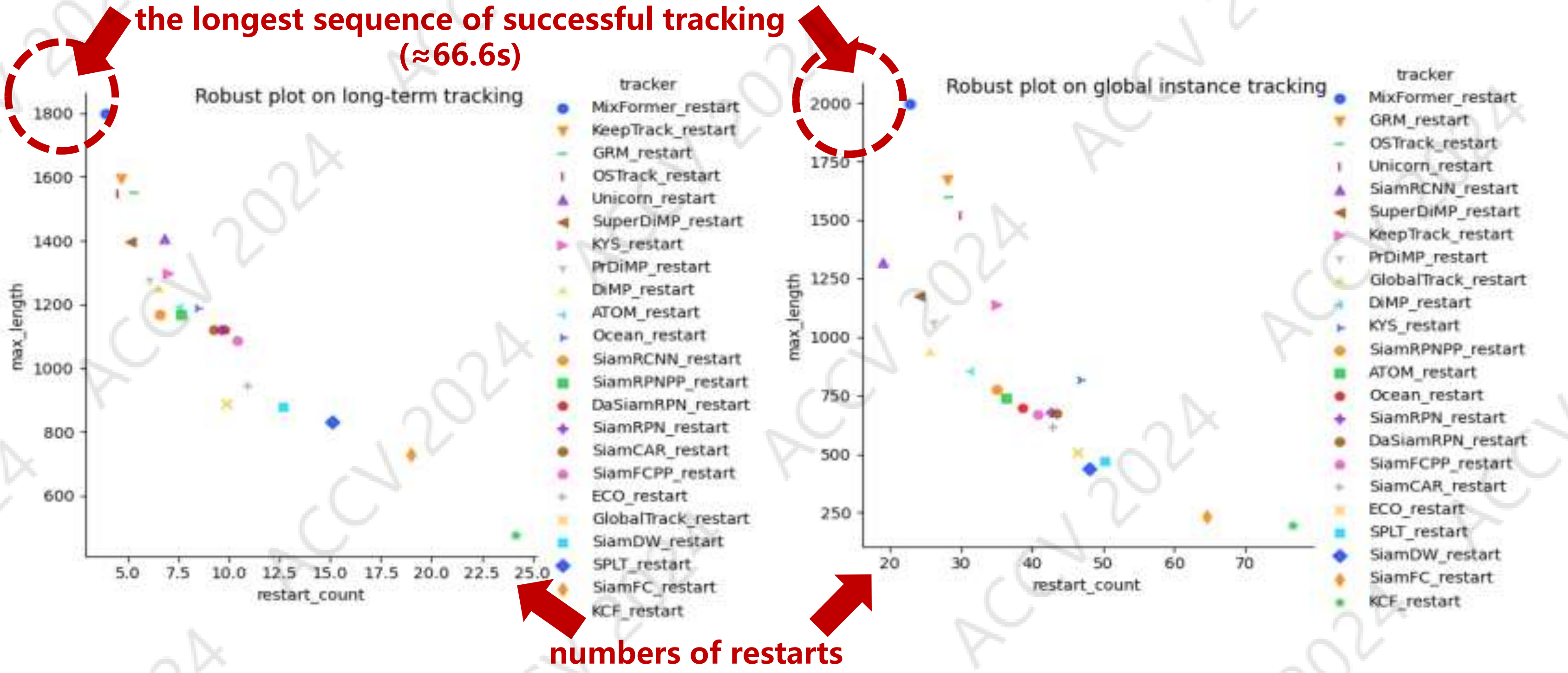
S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (IJCV), 2024.

Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**

➢ **Some key issues that are masked by traditional evaluation methods:** How well does the algorithm track over long sequences?

  - **Restart mechanism (R-OPE):** When an algorithm failure is detected, the algorithm is reinitialized at the nearest restart point.
  - **Robust plot:** measures the number of restarts of the algorithm, and the longest sequence of successful tracking.



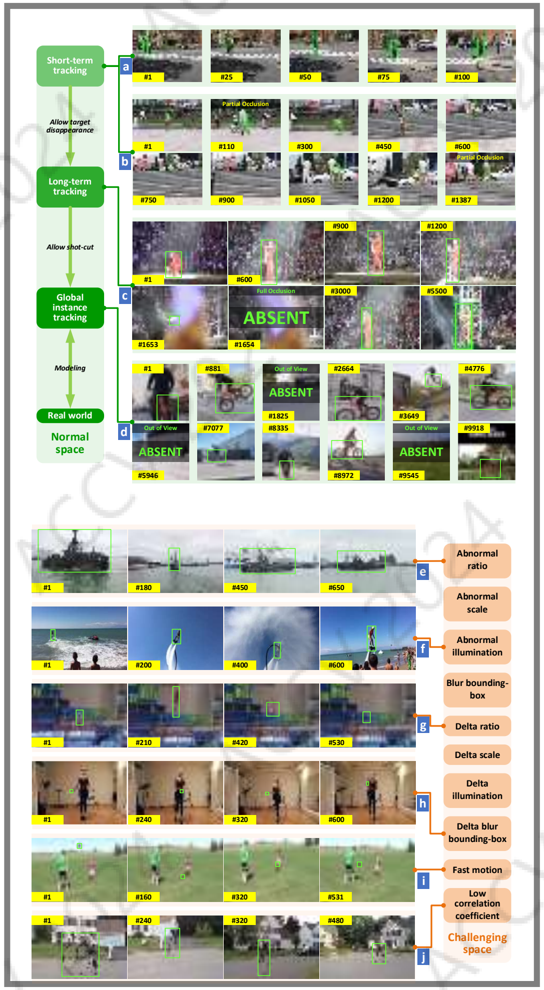
the longest sequence of successful tracking (≈66.6s)

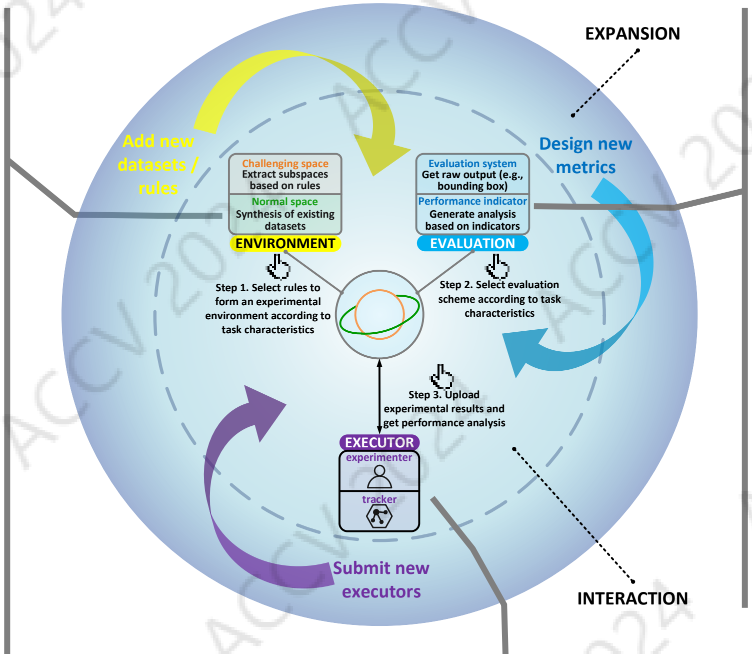
numbers of restarts

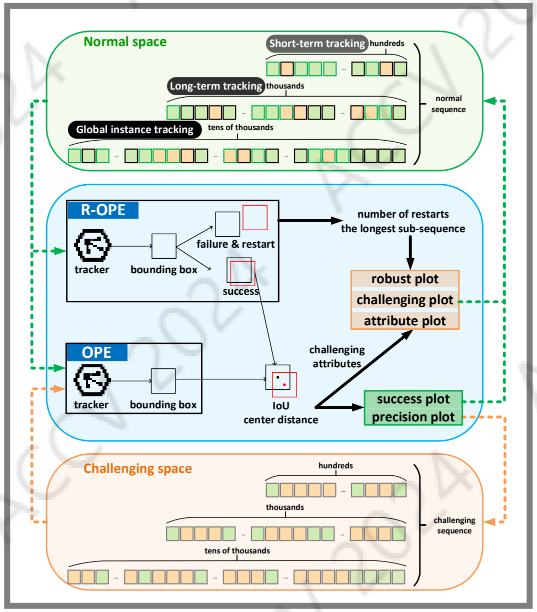
S. Hu, X. Zhao#, and K. Huang, "Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (IJCV), 2024.

Trend 2. More Realistic Data Environment

- **Example 1: SOTVerse (Dynamic and Open Task Space for VOT)**


Short-term tracking
Allow target disappearance
Long-term tracking
Allow shot-cut
Global instance tracking
Modeling
Real world
Normal space
a
b
c
d
Partial Occlusion
Full Occlusion
Out of View
ABSENT
e
f
g
h
i
j
Abnormal ratio
Abnormal scale
Abnormal illumination
Blur bounding-box
Delta ratio
Delta scale
Delta illumination
Delta blur bounding-box
Fast motion
Low correlation coefficient
Challenging space


EXPANSION
Add new datasets / rules
Design new metrics
Challenging space
Extract subspaces based on rules
Normal space
Synthesis of existing datasets
ENVIRONMENT
Evaluation system
Get raw output (e.g., bounding box)
Performance indicator
Generate analysis based on indicators
EVALUATION
Step 1. Select rules to form an experimental environment according to task characteristics
Step 2. Select evaluation scheme according to task characteristics
Step 3. Upload experimental results and get performance analysis
EXECUTOR
experimenter
tracker
Submit new executors
INTERACTION


Normal space
Short-term tracking
hundreds
Long-term tracking
thousands
Global instance tracking
tens of thousands
normal sequence
R-OPE
tracker
bounding box
failure & restart
success
number of restarts
the longest sub-sequence
robust plot
challenging plot
attribute plot
challenging attributes
OPE
tracker
bounding box
IoU
center distance
success plot
precision plot
Challenging space
hundreds
thousands
tens of thousands
challenging sequence

- ➢ Adapt to different experimental environments through multiple evaluation methods
- ➢ Comprehensive analysis of algorithms' bottlenecks

S. Hu, X. Zhao#, and K. Huang, *"Sotverse: A user-defined task space of single object tracking," International Journal of Computer Vision (**IJCV**), 2024.*

ACCV 2024

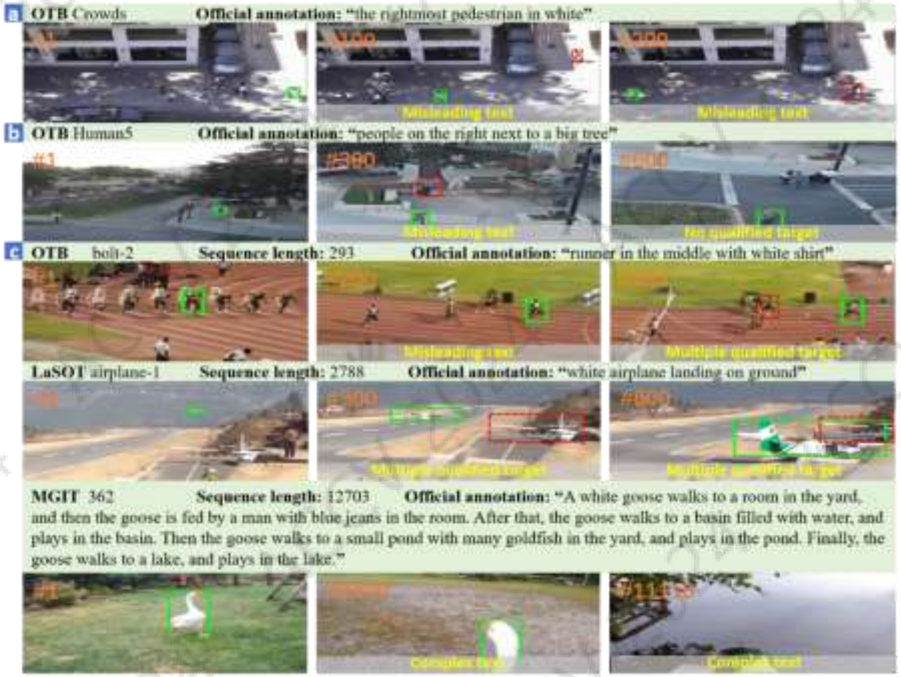
Trend 2. More Realistic Data Environment

- **Example 2: DTVLT (Diverse Multimodal Benchmark for VLT)**

➢ **Motivation:** Most VLT benchmarks are annotated in a **single granularity** and **lack a coherent semantic framework** to provide scientific guidance.

➢ Current VLT benchmarks considers studying from different perspective :

  - **Limitations 1.** Semantic annotations in OTB99_Lang mainly describe **the first frame**, which may **misguide the algorithm**.
  - **Limitations 2.** Sequence in MGIT has such **complex text** that they are **not conducive to algorithmic learning**.


a
b
c
#1
#100
#200
#1
#380
Misleading text
Misleading text
Misleading text
No qualified target
Misleading text
Multiple qualified target
Multiple qualified target
Multiple qualified target
#11
Complex text
Complex text

OTB Crowds **Official annotation:** "the rightmost pedestrian in white"

OTB Human5 **Official annotation:** "people on the right next to a big tree"

OTB bolt-2 **Sequence length:** 293 **Official annotation:** "runner in the middle with white shirt"

LaSOT airplane-1 **Sequence length:** 2788 **Official annotation:** "white airplane landing on ground"

MGIT 362 **Sequence length:** 12703 **Official annotation:** "A white goose walks to a room in the yard, and then the goose is fed by a man with blue jeans in the room. After that, the goose walks to a basin filled with water, and plays in the basin. Then the goose walks to a small pond with many goldfish in the yard, and plays in the pond. Finally, the goose walks to a lake, and plays in the lake."

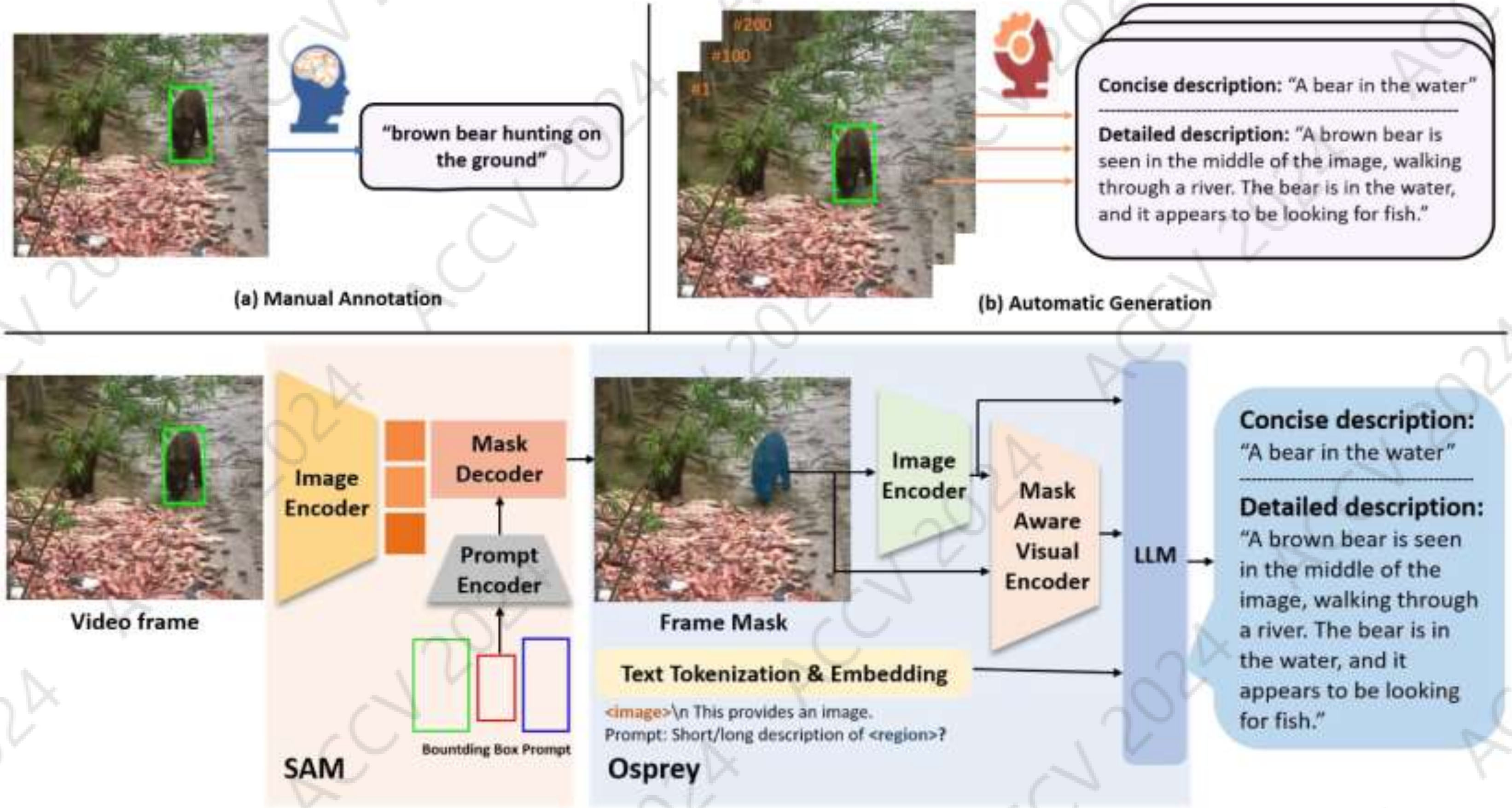
➢ **Research objective :** Using **LLM** to provide **multi-granularity semantic information** for VLT from **efficient and diverse** perspectives, enabling fine-grained evaluation. This work can be extended to more datasets to support vision datasets understanding.

Li X, Hu S, Feng X, et al. DTVLT: A Multi-modal Diverse Text Benchmark for Visual Language Tracking Based on LLM[J]. arXiv preprint arXiv:2410.02492, 2024.

ACCV 2024

Trend 2. More Realistic Data Environment

- **Example 2: DTVLT (Diverse Multimodal Benchmark for VLT)**
  - ➢ **Diverse texts** matter → Integrating the **LLM** into the text generation process, offer a **diverse environment** conducive to VLT research.


#200
#100
#1
Video frame
Image Encoder
Mask Decoder
Prompt Encoder
Bountding Box Prompt
SAM
Frame Mask
Image Encoder
Mask Aware Visual Encoder
LLM
Text Tokenization & Embedding
Osprey

"brown bear hunting on the ground"

(a) Manual Annotation

Concise description: "A bear in the water"

Detailed description: "A brown bear is seen in the middle of the image, walking through a river. The bear is in the water, and it appears to be looking for fish."

(b) Automatic Generation

<image>\n This provides an image.
Prompt: Short/long description of <region>?

Concise description: "A bear in the water"

Detailed description: "A brown bear is seen in the middle of the image, walking through a river. The bear is in the water, and it appears to be looking for fish."

(c) Framework of DTLLM-VLT

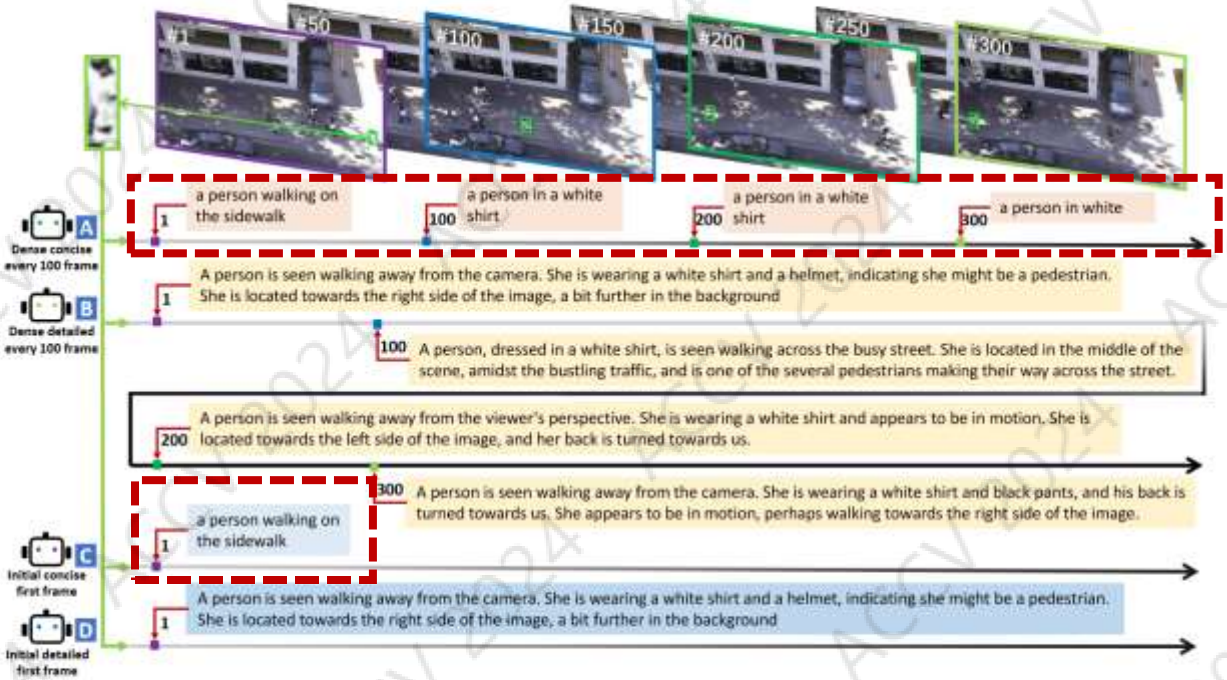
Li X, Hu S, Feng X, et al. DTVLT: A Multi-modal Diverse Text Benchmark for Visual Language Tracking Based on LLM[J]. arXiv preprint arXiv:2410.02492, 2024.

Trend 2. More Realistic Data Environment

- **Example 2: DTVLT (Diverse Multimodal Benchmark for VLT)**

➢ **Applying multi-granularity generation**

  - **Initial texts**: Following the text annotations method in OTB99_Lang and TNL2K, we generate text for the **initial frame** of each video.
  - **Dense texts**: Considering the worst situation and infer that the algorithm lacks an efficient memory system. Consequently, at 25 FPS, equating to **every 100 frames** in 4 seconds, we supply the algorithm with relevant generated text.


#1
#50
#100
#150
#200
#250
#300
A
Dense concise every 100 frame
B
Dense detailed every 100 frame
C
Initial concise first frame
D
Initial detailed first frame

1 a person walking on the sidewalk

100 a person in a white shirt

200 a person in a white shirt

300 a person in white

1 A person is seen walking away from the camera. She is wearing a white shirt and a helmet, indicating she might be a pedestrian. She is located towards the right side of the image, a bit further in the background

100 A person, dressed in a white shirt, is seen walking across the busy street. She is located in the middle of the scene, amidst the bustling traffic, and is one of the several pedestrians making their way across the street.

200 A person is seen walking away from the viewer's perspective. She is wearing a white shirt and appears to be in motion. She is located towards the left side of the image, and her back is turned towards us.

300 A person is seen walking away from the camera. She is wearing a white shirt and black pants, and his back is turned towards us. She appears to be in motion, perhaps walking towards the right side of the image.

1 a person walking on the sidewalk

1 A person is seen walking away from the camera. She is wearing a white shirt and a helmet, indicating she might be a pedestrian. She is located towards the right side of the image, a bit further in the background

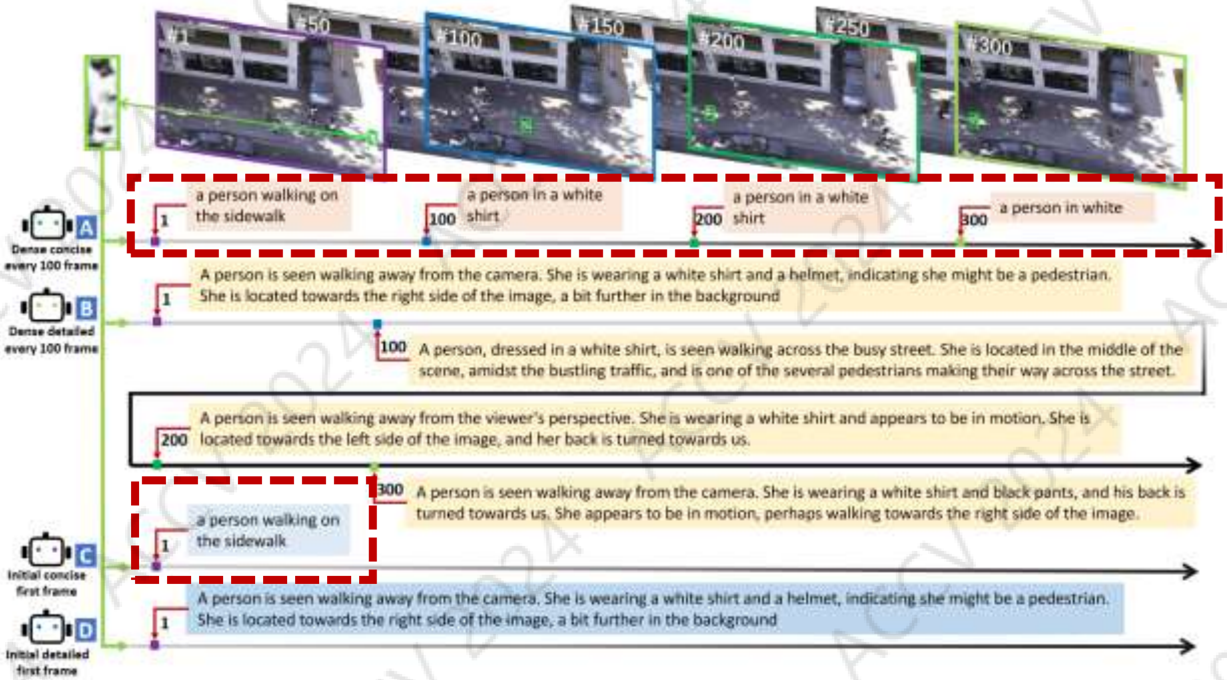
Li X, Hu S, Feng X, et al. DTVLT: A Multi-modal Diverse Text Benchmark for Visual Language Tracking Based on LLM[J]. arXiv preprint arXiv:2410.02492, 2024.

Trend 2. More Realistic Data Environment

- Example 2: DTVLT (Diverse Multimodal Benchmark for VLT)

➢**Applying multi-granularity generation**

- **Concise texts:** If the BBox already sufficiently describes the temporal and spatial changes of the object, the text descriptions should focus on providing **essential semantic details** like the category and positions of the object.
- **Dense texts**: In cases where the BBox lacks sufficient information for effective learning by the tracker, more **elaborate texts are necessary** to compensate for the missing temporal and spatial relationships.


#1
#50
#100
#150
#200
#250
#300
A
Dense concise every 100 frame
B
Dense detailed every 100 frame
C
Initial concise first frame
D
Initial detailed first frame
1
100
200
300
1
100
200
300
1
1

a person walking on the sidewalk

a person in a white shirt

a person in a white shirt

a person in white

A person is seen walking away from the camera. She is wearing a white shirt and a helmet, indicating she might be a pedestrian. She is located towards the right side of the image, a bit further in the background

A person, dressed in a white shirt, is seen walking across the busy street. She is located in the middle of the scene, amidst the bustling traffic, and is one of the several pedestrians making their way across the street.

A person is seen walking away from the viewer's perspective. She is wearing a white shirt and appears to be in motion. She is located towards the left side of the image, and her back is turned towards us.

A person is seen walking away from the camera. She is wearing a white shirt and black pants, and his back is turned towards us. She appears to be in motion, perhaps walking towards the right side of the image.

a person walking on the sidewalk

A person is seen walking away from the camera. She is wearing a white shirt and a helmet, indicating she might be a pedestrian. She is located towards the right side of the image, a bit further in the background

Li X, Hu S, Feng X, et al. DTVLT: A Multi-modal Diverse Text Benchmark for Visual Language Tracking Based on LLM[J]. arXiv preprint arXiv:2410.02492, 2024.

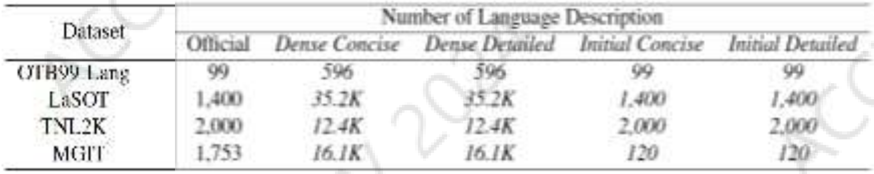
ACCV 2024

Trend 2. More Realistic Data Environment

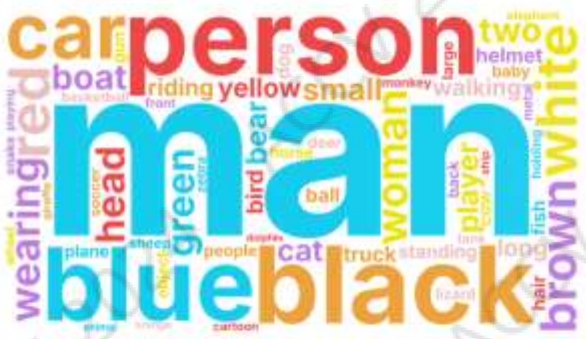
- **Example 2: DTVLT (Diverse Multimodal Benchmark for VLT)**

➢ **Diverse Generation**

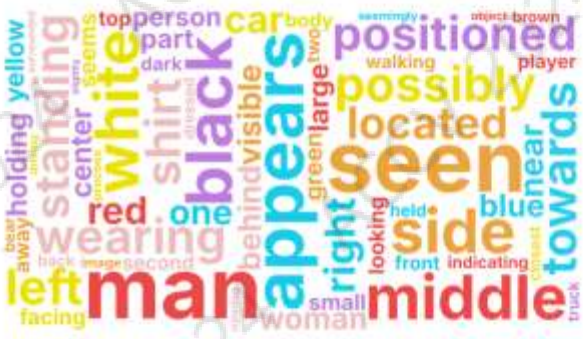
  - **1.9M** words
  - **14.8K** non-repetitive words.
  - 7,238 initial descriptions
  - 128.4K dense descriptions



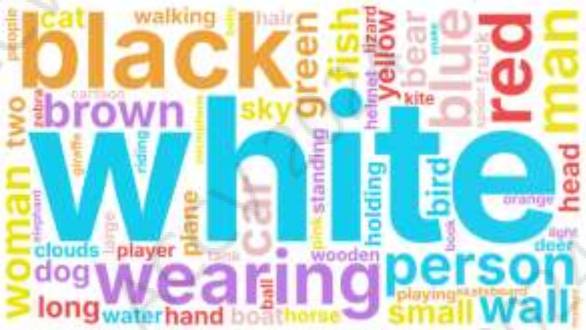
Dataset	Number of Language Description				
	Official	*Dense Concise*	*Dense Detailed*	*Initial Concise*	*Initial Detailed*
OTB99 Lang	99	596	596	99	99
LaSOT	1,400	*35.2K*	*35.2K*	*1,400*	*1,400*
TNL2K	2,000	*12.4K*	*12.4K*	*2,000*	*2,000*
MGIT	1,753	*16.1K*	*16.1K*	*120*	*120*



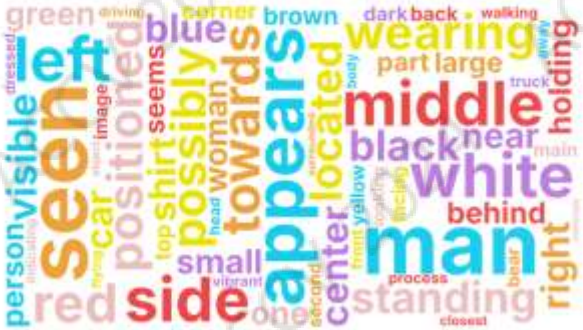
(a) The word cloud of initial concise texts



(b) The word cloud of initial detailed texts



(c) The word cloud of dense concise texts



(d) The word cloud of dense detailed texts

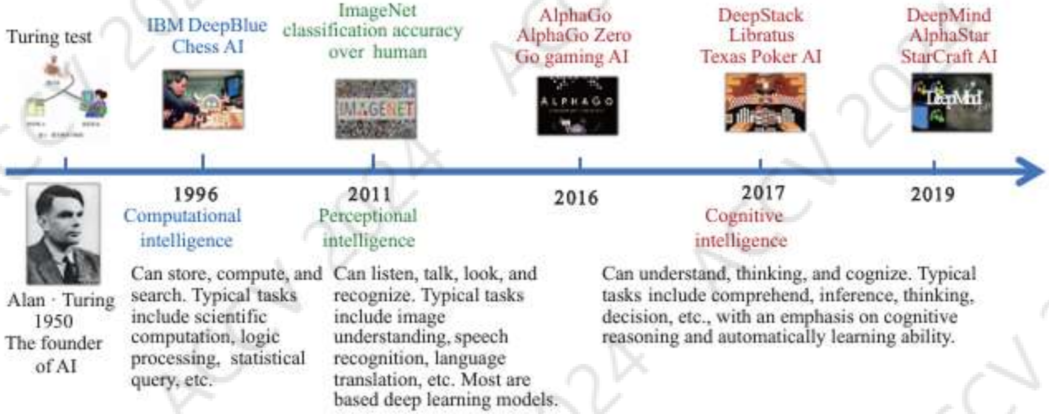
Li X, Hu S, Feng X, et al. DTVLT: A Multi-modal Diverse Text Benchmark for Visual Language Tracking Based on LLM[J]. arXiv preprint arXiv:2410.02492, 2024.

Trend 3. More Human-like Executors

ACCV 2024

Trend 3. More Human-like Executors

- **Optimizing algorithms from the perspective of human-like modeling: understanding video content more like humans.**


Turing test
IBM DeepBlue Chess AI
ImageNet classification accuracy over human
AlphaGo AlphaGo Zero Go gaming AI
DeepStack Libratus Texas Poker AI
DeepMind AlphaStar StarCraft AI
IMAGENET
ALPHAGO
DeepMind
1996
Computational intelligence
2011
Perceptional intelligence
2016
2017
Cognitive intelligence
2019
Alan · Turing
1950
The founder of AI

Can store, compute, and search. Typical tasks include scientific computation, logic processing, statistical query, etc.

Can listen, talk, look, and recognize. Typical tasks include image understanding, speech recognition, language translation, etc. Most are based deep learning models.

Can understand, thinking, and cognize. Typical tasks include comprehend, inference, thinking, decision, etc., with an emphasis on cognitive reasoning and automatically learning ability.

➢ **Goal of Computer Vision Research:** Equip machines with human-like visual intelligence. The goal is to achieve machine intelligence at multiple levels**.**

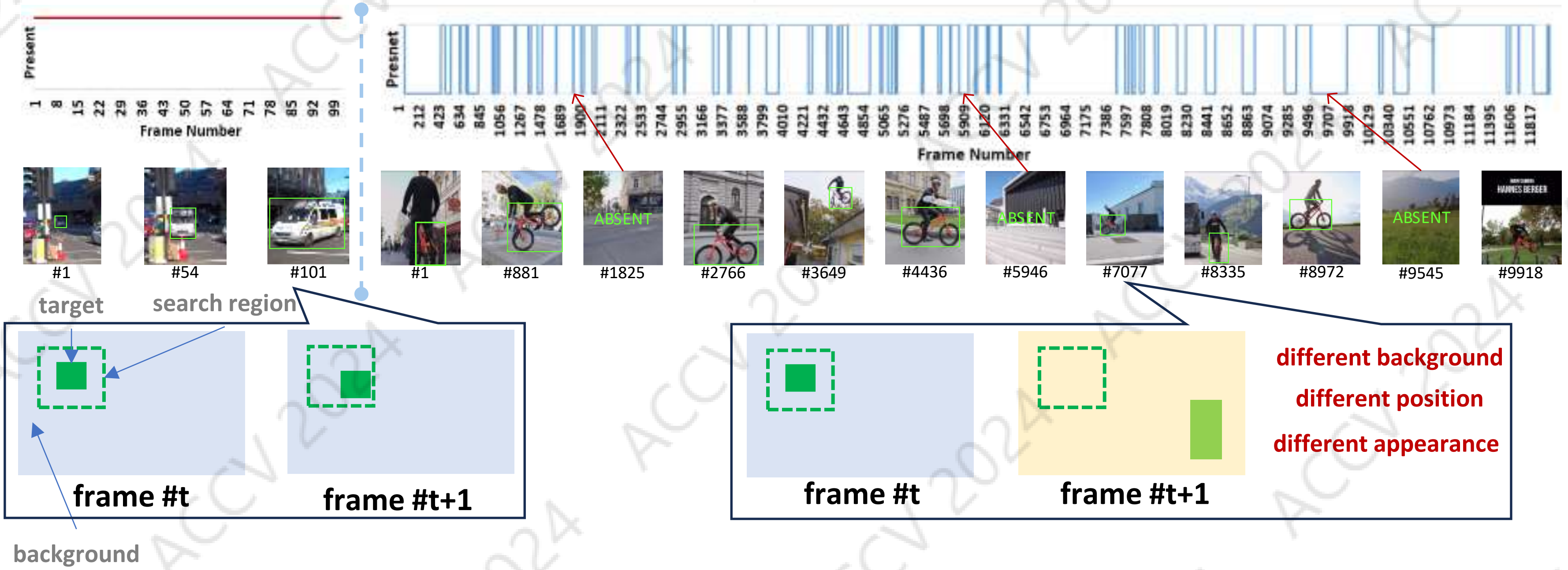
- **Computational Intelligence:** Responsible for signal processing, logical processing, and statistical calculations, serving as the foundation for higher intelligence levels.
- **Perceptual Intelligence:** Involves the ability to perceive and capture visual information from the environment, such as image recognition and object detection.
- **Cognitive Intelligence:** Includes memory, prediction, and reasoning capabilities, forming the basis for understanding and inferring future actions.

Huang K Q, Xing J L, Zhang J G, et al. Intelligent technologies of human-computer gaming (in Chinese). Sci Sin Inform, 2020, 50: 540–550, doi: 10.1360/N112019-00048

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**

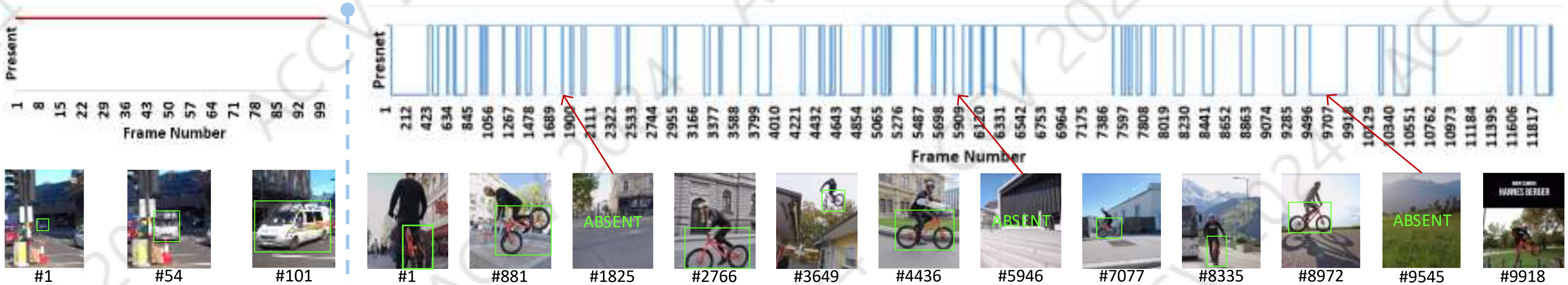

Present
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Frame Number
Presnet
1 212 423 634 845 1056 1267 1478 1689 1900 2111 2322 2533 2744 2955 3166 3377 3588 3799 4010 4221 4432 4643 4854 5065 5276 5487 5698 5909 6120 6331 6542 6753 6964 7175 7386 7597 7808 8019 8230 8441 8652 8863 9074 9285 9496 9707 9918 10129 10340 10551 10762 10973 11184 11395 11606 11817
Frame Number
#1
#54
#101
#1
#881
ABSENT
#1825
#2766
#3649
#4436
ABSENT
#5946
#7077
#8335
#8972
ABSENT
#9545
#9918
target
search region
frame #t
frame #t+1
background
frame #t
frame #t+1
different background
different position
different appearance

local search: applies **only to continuous motion assumption**

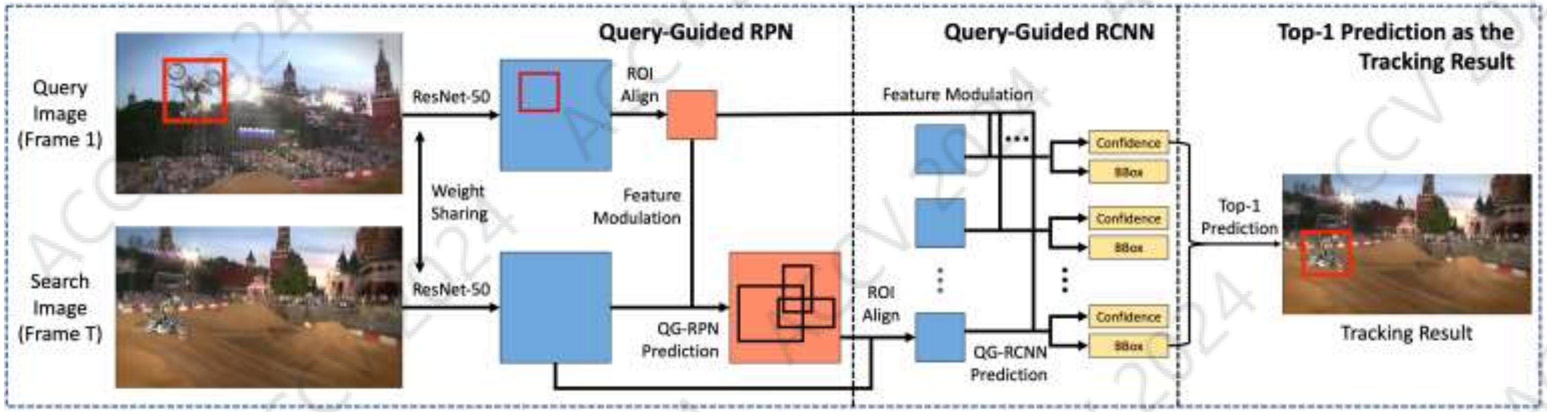
ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**



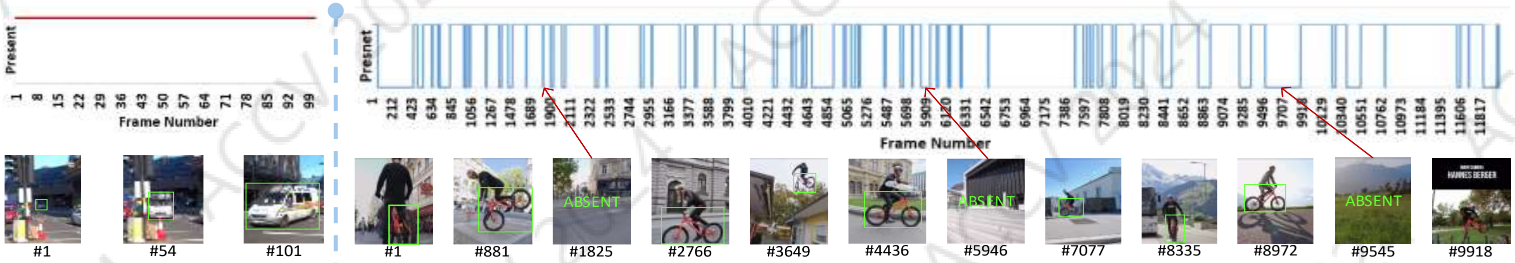
- **local search:** applies **only to continuous motion assumption**


Query-Guided RPN
Query-Guided RCNN
Top-1 Prediction as the Tracking Result
Query Image (Frame 1)
Search Image (Frame T)
ResNet-50
ResNet-50
Weight Sharing
ROI Align
Feature Modulation
QG-RPN Prediction
Feature Modulation
ROI Align
Confidence
BBox
Confidence
BBox
Confidence
BBox
QG-RCNN Prediction
Top-1 Prediction
Tracking Result

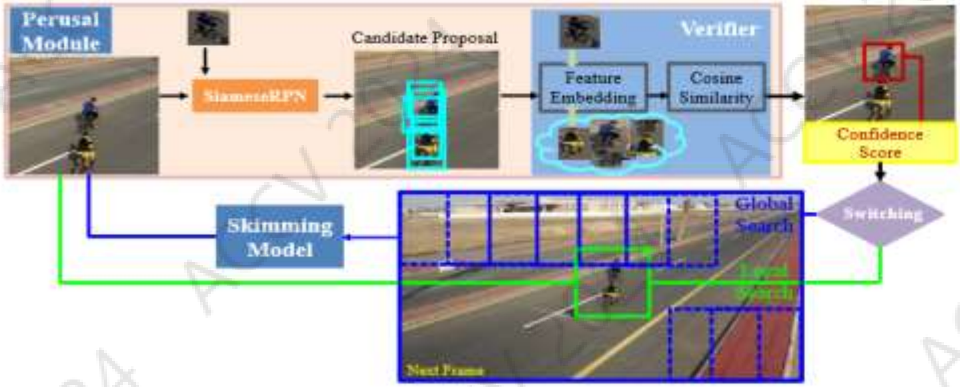
- **global search:** zero cumulative error, but it is **slow** and **easily interfered by background**

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**



- **local search:** applies **only to continuous motion assumption**
- **global search:** zero cumulative error, but it is **slow** and **easily interfered by background**


Perusal Module
Candidate Proposal
SiameseRPN
Verifier
Feature Embedding
Cosine Similarity
Confidence Score
Switching
Skimming Model
Global Search
Local Search
Next Frame

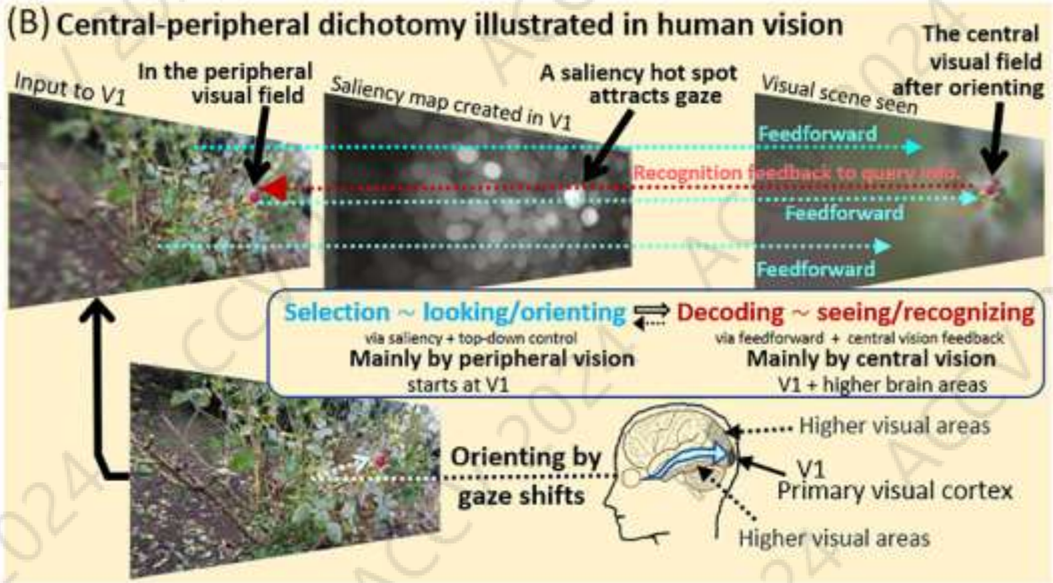
- **local search + global search:** good idea, but **the timing of the switch is difficult to determine**

How the human visual system accurately finds the target in a new frame?

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**


(B) Central-peripheral dichotomy illustrated in human vision
Input to V1
In the peripheral visual field
Saliency map created in V1
A saliency hot spot attracts gaze
Visual scene seen
The central visual field after orienting
Feedforward
Recognition feedback to query info.
Feedforward
Feedforward
Selection ~ looking/orienting
via saliency + top-down control
Mainly by peripheral vision
starts at V1
Decoding ~ seeing/recognizing
via feedforward + central vision feedback
Mainly by central vision
V1 + higher brain areas
Orienting by gaze shifts
Higher visual areas
V1
Primary visual cortex
Higher visual areas

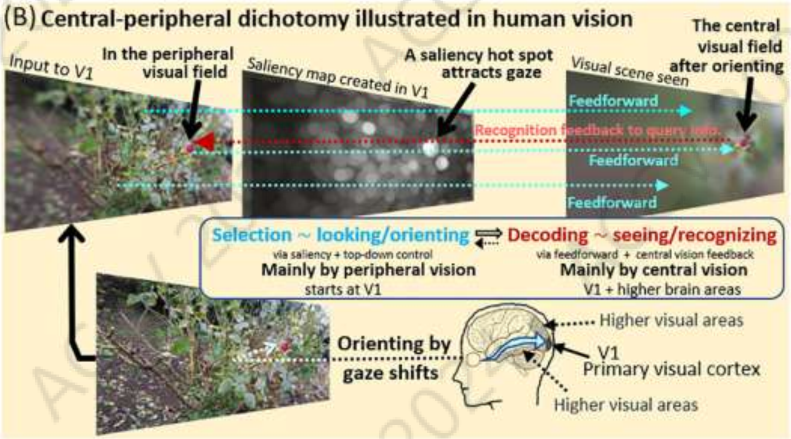
➢ **Central-Peripheral Dichotomy:** The human visual system is divided into central vision and peripheral vision, both playing distinct roles in the process of **visual perception.**

- **Peripheral Vision:** Responsible for detecting a wide visual field, mainly used for identifying salient areas in the environment.
- **Central Vision:** Responsible for fine visual processing, mainly used for target recognition and decoding.

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**


(B) Central-peripheral dichotomy illustrated in human vision
Input to V1
In the peripheral visual field
Saliency map created in V1
A saliency hot spot attracts gaze
Visual scene seen
The central visual field after orienting
Feedforward
Recognition feedback to query info.
Feedforward
Feedforward
Selection ~ looking/orienting
via saliency + top-down control
Mainly by peripheral vision
starts at V1
Decoding ~ seeing/recognizing
via feedforward + central vision feedback
Mainly by central vision
V1 + higher brain areas
Orienting by gaze shifts
Higher visual areas
V1
Primary visual cortex
Higher visual areas

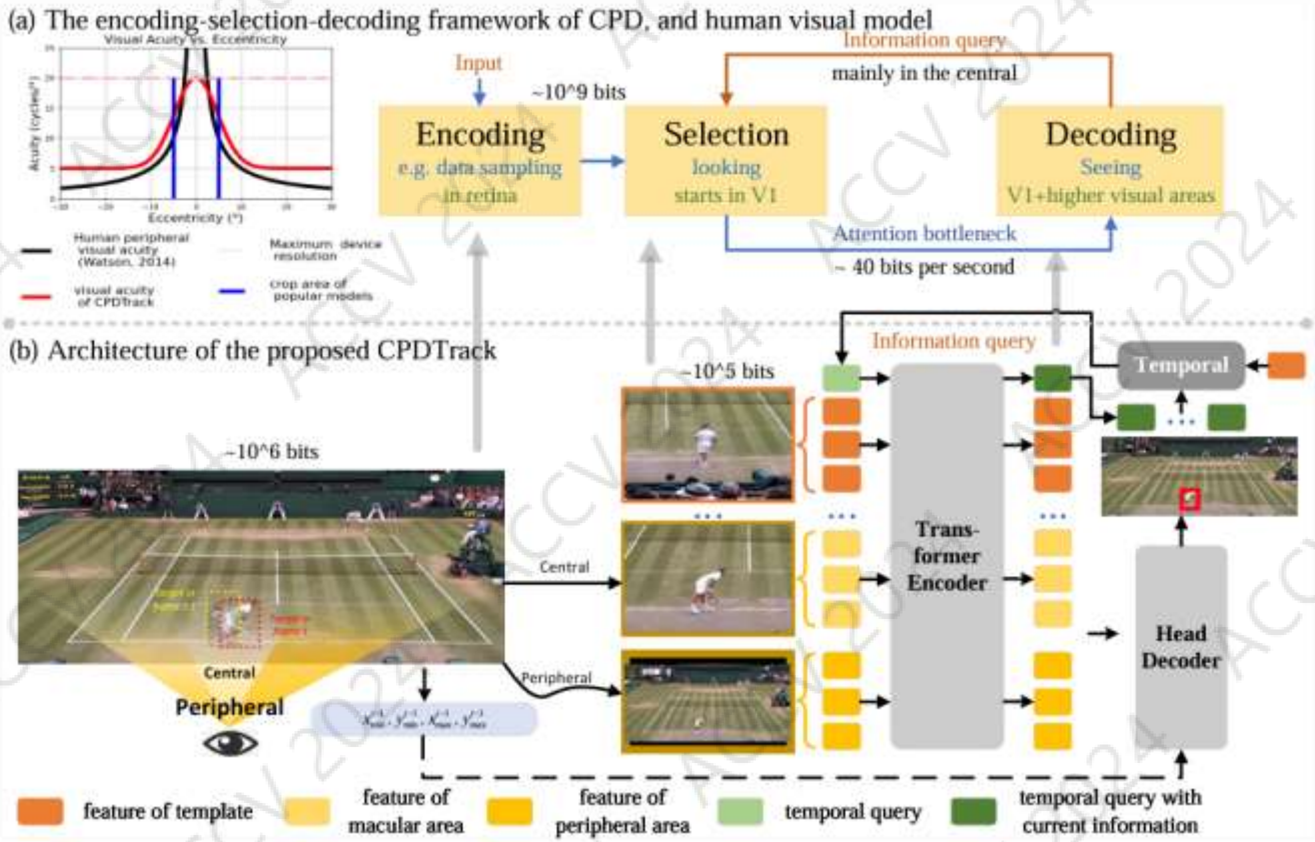
➢ **Process of Encoding, Selection, and Decoding:**
  - **Encoding:** Visual information from the peripheral field enters V1, generating a saliency hotspot map for further processing.
  - **Selection:** Peripheral vision uses saliency and top-down control to guide gaze shifts to areas of interest.
  - **Decoding:** Central vision decodes detailed information of the selected area through feedforward and feedback streams, enabling object recognition and scene understanding.

➢ Peripheral vision **scans the environment broadly** for potential targets, while central vision **focuses on high-precision visual decoding**.

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**


(a) The encoding-selection-decoding framework of CPD, and human visual model
Visual Acuity vs. Eccentricity
Acuity (cycles/°)
Eccentricity (°)
Human peripheral visual acuity (Watson, 2014)
Maximum device resolution
visual acuity of CPDTrack
crop area of popular models
Input
~10^9 bits
Information query
mainly in the central
Encoding
e.g. data sampling in retina
Selection
looking
starts in V1
Decoding
Seeing
V1+higher visual areas
Attention bottleneck
~ 40 bits per second
(b) Architecture of the proposed CPDTrack
~10^6 bits
~10^5 bits
Information query
Temporal
Central
Peripheral
Central
Peripheral
Trans-former Encoder
Head Decoder
feature of template
feature of macular area
feature of peripheral area
temporal query
temporal query with current information

- ➢ We constructed a model of the Central-Peripheral Dichotomy theory in cognitive science, utilizing the one-stream structure in visual object tracking.

D. Zhang, S. Hu, et al., "Beyond accuracy: Tracking more like human via visual search," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**

We study the discontinuity of the target state in space and time (i.e., *STDChallenge*), which comprises two challenges: **absent** and **shot-cut**.

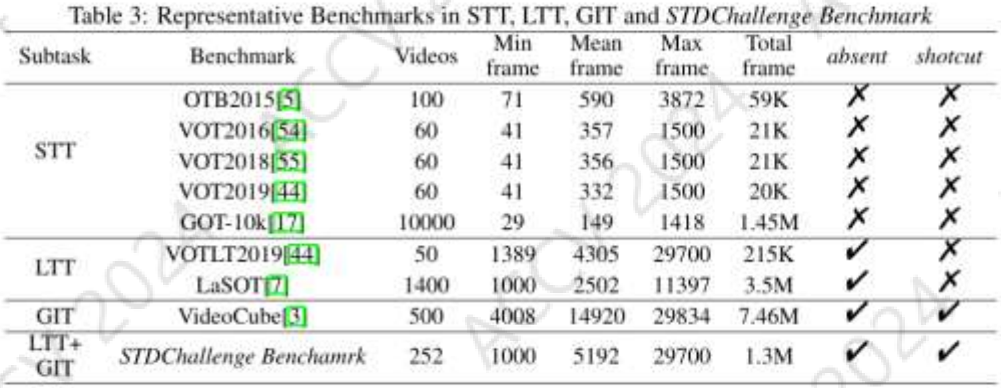


Table 3: Representative Benchmarks in STT, LTT, GIT and *STDChallenge Benchmark*

Subtask	Benchmark	Videos	Min frame	Mean frame	Max frame	Total frame	*absent*	*shotcut*
STT	OTB2015[5]	100	71	590	3872	59K	✗	✗
	VOT2016[54]	60	41	357	1500	21K	✗	✗
	VOT2018[55]	60	41	356	1500	21K	✗	✗
	VOT2019[44]	60	41	332	1500	20K	✗	✗
	GOT-10k[17]	10000	29	149	1418	1.45M	✗	✗
LTT	VOTLT2019[44]	50	1389	4305	29700	215K	✓	✗
	LaSOT[7]	1400	1000	2502	11397	3.5M	✓	✗
GIT	VideoCube[3]	500	4008	14920	29834	7.46M	✓	✓
LTT+ GIT	*STDChallenge Benchamrk*	252	1000	5192	29700	1.3M	✓	✓

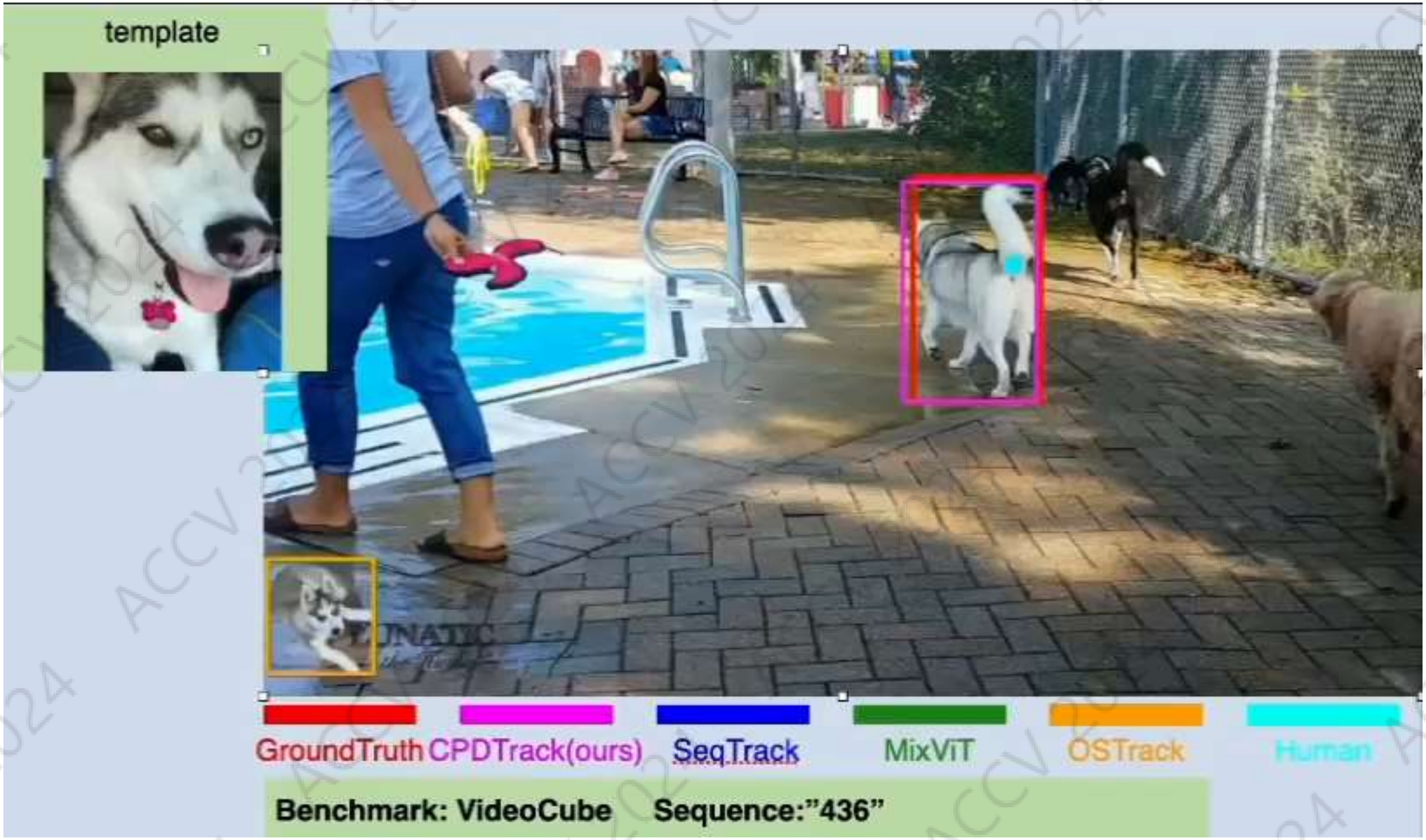
$$STD = \frac{(n_a + n_s) \cdot l_a}{l^2},$$

- ➢ We extracted sequences from LTT and GIT tasks that include the STDChallenge to form the **STDChallenge Benchmark**, aiming to suppress the bias of a single dataset.
- ➢ At the same time, we **quantified the difficulty of the STDChallenge**, taking into account the challenges of 'disappearance-reappearance' and 'shot switching' within the sequences.
- ➢ We divided the STDChallenge Benchmark into **three groups with different difficulty levels** based on the STD metric and selected five sequences from each group to form the STDChallenge Turing, which is used for the **Visual Turing Test.**

D. Zhang, S. Hu, et al., "Beyond accuracy: Tracking more like human via visual search," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**

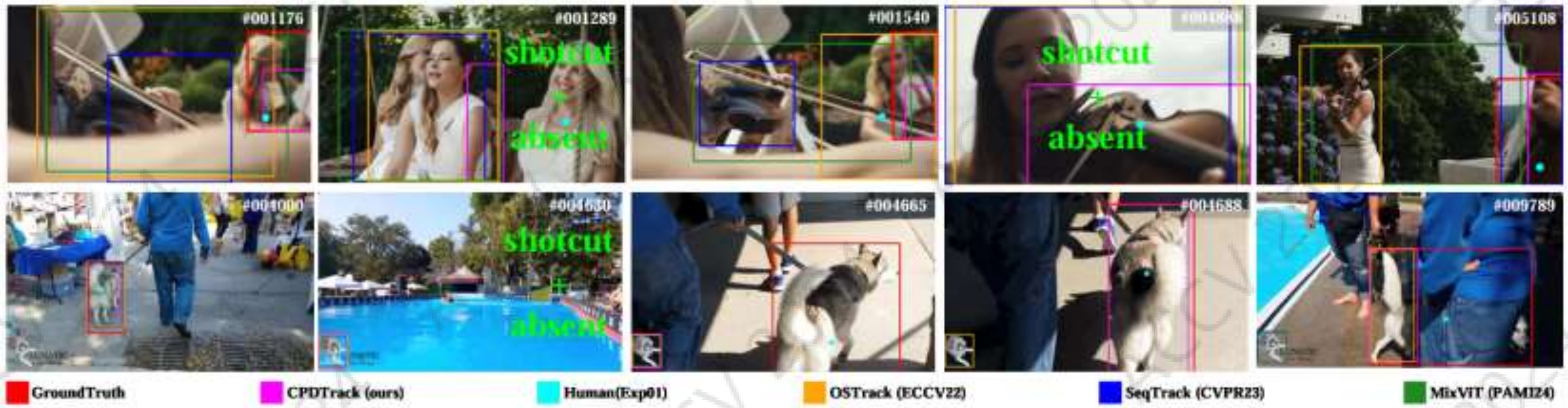

template
GroundTruth
CPDTrack(ours)
SeqTrack
MixViT
OSTrack
Human
Benchmark: VideoCube
Sequence:"436"

D. Zhang, **S. Hu**, et al., *"Beyond accuracy: Tracking more like human via visual search," in the 38th Conference on Neural Information Processing Systems (NeurIPS).*

ACCV 2024

Trend 3. More Human-like Executors

- **Example 1. Human-like VOT via Visual Search Ability (Better Perceptual Intelligence)**



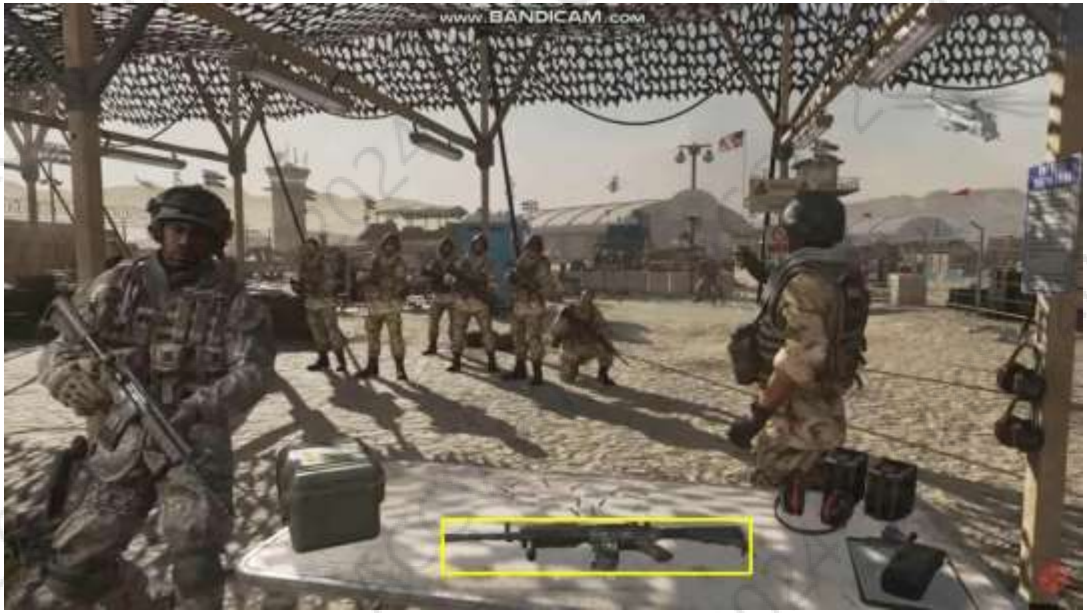
#001176
#001289
#001540
#005108
#004665
#001688
#009789
shotcut
absent
🟥 GroundTruth
🟪 CPDTrack (ours)
🟦 Human(Exp01)
🟧 OSTrack (ECCV22)
🟦 SeqTrack (CVPR23)
🟩 MixViT (PAMI24)

- ➢ Human results do not necessarily mean correctness, but **humans can usually quickly re-locate the target** after the STDChallenge.
- ➢ In the second image of the first row, **humans can recognize environmental factors closely related to the target.**
- ➢ In the second image of the second row, even when the target is absent, **humans are not distracted by the background.**
- ➢ In the fifth image, **humans are robust to occlusion**.

D. Zhang, S. Hu, et al., "Beyond accuracy: Tracking more like human via visual search," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

Trend 3. More Human-like Executors

- **Example 2. Human-like VLT via Memory Modeling (Better Cognitive Intelligence)**



"The gun on the table"

➢ **Limitations of Static Cues:**
  - Text-template cues are **static and fixed**, whereas objects in the video are **dynamically changing**.
  - Static cues **cannot continuously provide reliable reference** for similarity matching.

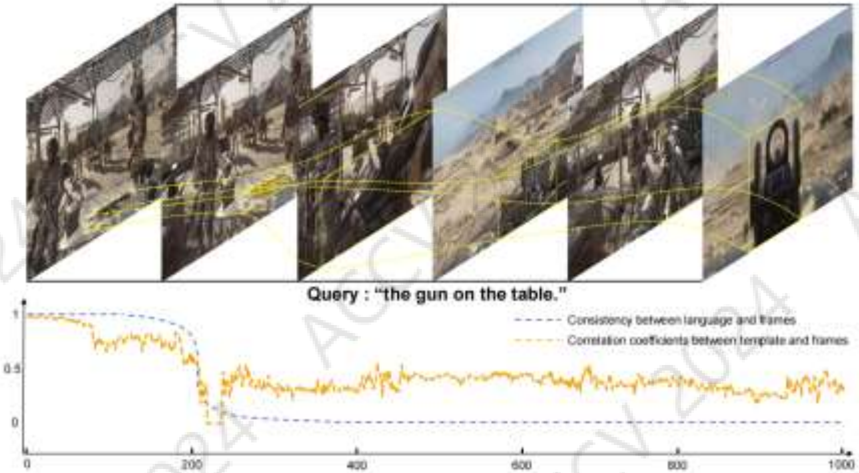
X. Feng, X. Li, S. Hu, et al., "Memvlt: Visual-language tracking with adaptive memory-based prompts," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

ACCV 2024

Trend 3. More Human-like Executors

- **Example 2. Human-like VLT via Memory Modeling (Better Cognitive Intelligence)**

Template



➢ **Necessity of Introducing Temporal Information:**
  - **Dynamically Changing Targets:** The described scene in the text may **not align** with the actual target in the video, and the target **undergoes significant appearance changes** across frames, leading to lower matching accuracy with the image template.
  - **Changing Environmental Factors**: In the video sequence, both the background and the state of the target are constantly changing, making it **difficult to handle these dynamics with static templates alone**.

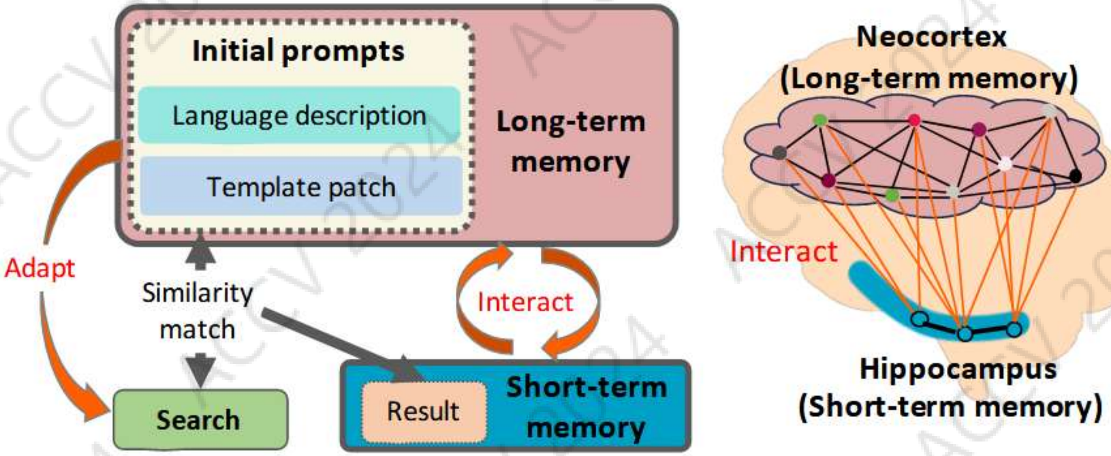
➢ **Utilize Temporal Information to Provide Dynamic Cues:** By introducing temporal information, the tracking task can make use of frame-to-frame changes, enabling better target localization and tracking.

X. Feng, X. Li, S. Hu, et al., "Memvlt: Visual-language tracking with adaptive memory-based prompts," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

ACCV 2024

Trend 3. More Human-like Executors

- **Example 2. Human-like VLT via Memory Modeling (Better Cognitive Intelligence)**


Initial prompts
Language description
Template patch
Long-term memory
Adapt
Similarity match
Search
Interact
Result
Short-term memory
Neocortex
(Long-term memory)
Interact
Hippocampus
(Short-term memory)

➢ **MemVLT:**

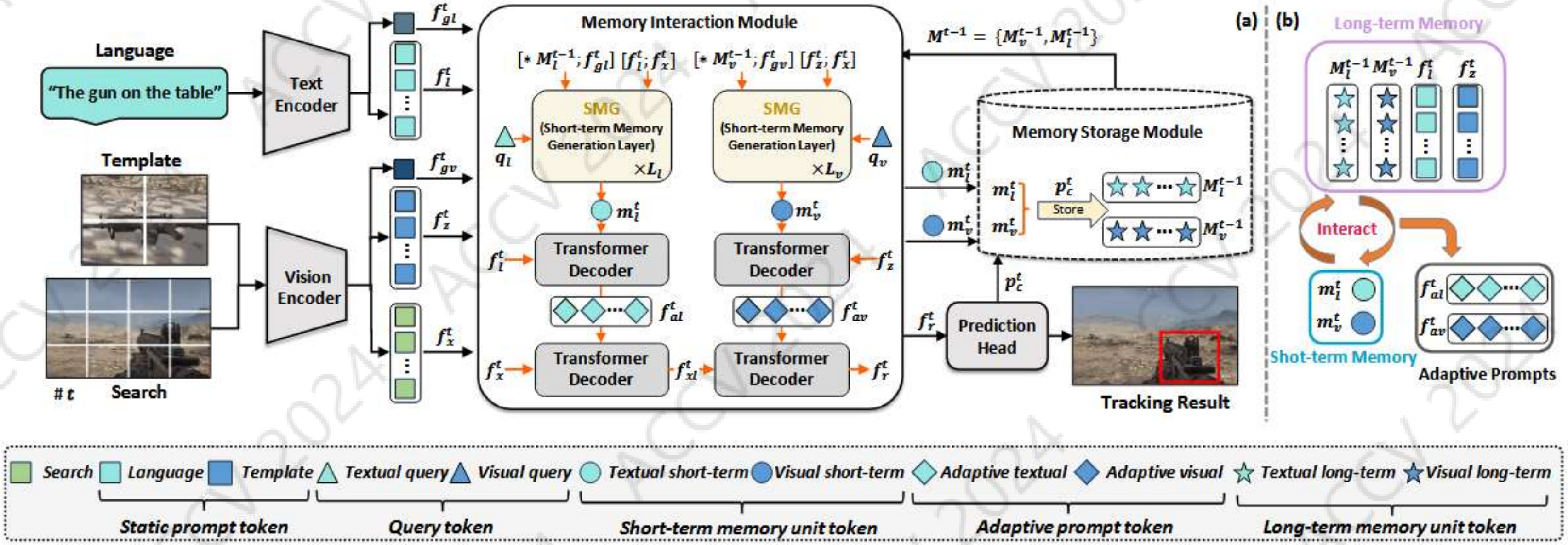
- Aims to address the issue where static, fixed multimodal cues struggle to continuously guide tracking of dynamically changing targets.
- Based on **complementary learning theory**, it models and stores dynamic changes in the target and adjusts the static template accordingly.
  - ☐ The human brain has two areas for storing memories: the **hippocampus for short-term memory** and the **neocortex for long-term memory.**
  - ☐ The **interaction between long and short-term memory** promotes human adaptation to different environments.

X. Feng, X. Li, S. Hu, et al., "Memvlt: Visual-language tracking with adaptive memory-based prompts," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

ACCV 2024

Trend 3. More Human-like Executors

- **Example 2. Human-like VLT via Memory Modeling (Better Cognitive Intelligence)**


Language
"The gun on the table"
Text Encoder
Template
Vision Encoder
# t
Search
f_{gl}^t
f_l^t
f_{gv}^t
f_z^t
f_x^t
Memory Interaction Module
$[* M_l^{t-1}; f_{gl}^t]$ $[f_l^t; f_x^t]$
$[* M_v^{t-1}; f_{gv}^t]$ $[f_z^t; f_x^t]$
SMG (Short-term Memory Generation Layer) $\times L_l$
SMG (Short-term Memory Generation Layer) $\times L_v$
q_l
q_v
m_l^t
m_v^t
Transformer Decoder
Transformer Decoder
f_l^t
f_z^t
f_{al}^t
f_{av}^t
Transformer Decoder
Transformer Decoder
f_x^t
f_{xl}^t
f_r^t
$M^{t-1} = \{M_v^{t-1}, M_l^{t-1}\}$
Memory Storage Module
m_l^t
m_v^t
p_c^t
Store
M_l^{t-1}
M_v^{t-1}
p_c^t
f_r^t
Prediction Head
Tracking Result
(a)
(b)
Long-term Memory
M_l^{t-1} M_v^{t-1} f_l^t f_z^t
Interact
m_l^t
m_v^t
Shot-term Memory
f_{al}^t
f_{av}^t
Adaptive Prompts
Search
Language
Template
Textual query
Visual query
Textual short-term
Visual short-term
Adaptive textual
Adaptive visual
Textual long-term
Visual long-term
Static prompt token
Query token
Short-term memory unit token
Adaptive prompt token
Long-term memory unit token

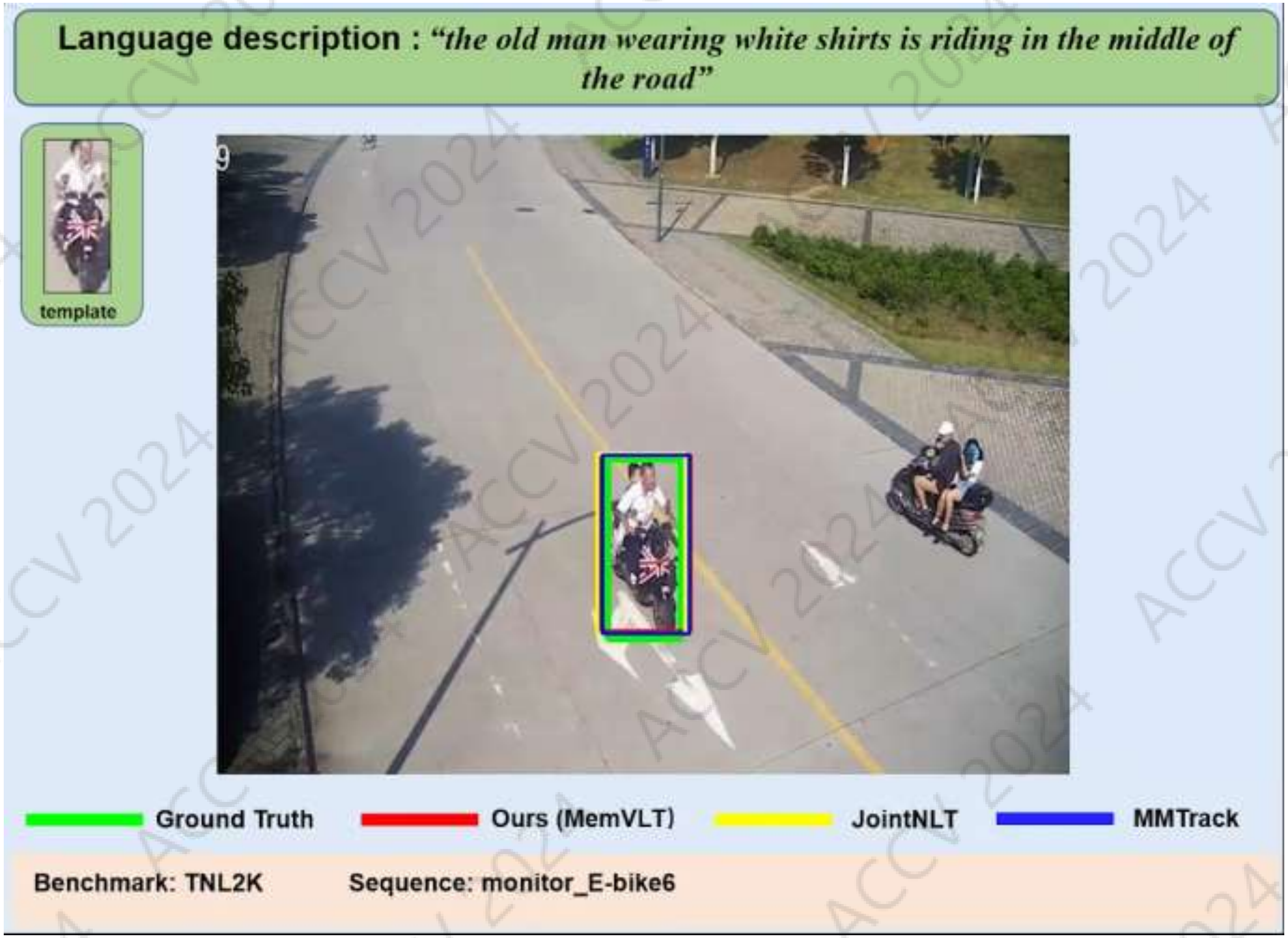
- **Core Design:**
  - **Memory Interaction Module:** Models dynamic changes in the target and adjusts the static template.
  - **Memory Storage Module:** Stores the dynamic features of the target.

X. Feng, X. Li, S. Hu, et al., "Memvlt: Visual-language tracking with adaptive memory-based prompts," in the 38th Conference on Neural Information Processing Systems (NeurIPS).

ACCV 2024

Trend 3. More Human-like Executors

- **Example 2. Human-like VLT via Memory Modeling (Better Cognitive Intelligence)**



Language description : *"the old man wearing white shirts is riding in the middle of the road"*

template
9
🟩 Ground Truth
🟥 Ours (MemVLT)
🟨 JointNLT
🟦 MMTrack

Benchmark: TNL2K Sequence: monitor_E-bike6

X. Feng, X. Li, S. Hu, et al., *"Memvlt: Visual-language tracking with adaptive memory-based prompts," in the 38th Conference on Neural Information Processing Systems (NeurIPS).*

ACCV 2024

Trend 4. More Intelligent Evaluation

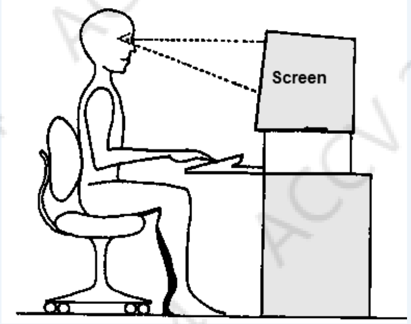
ACCV 2024

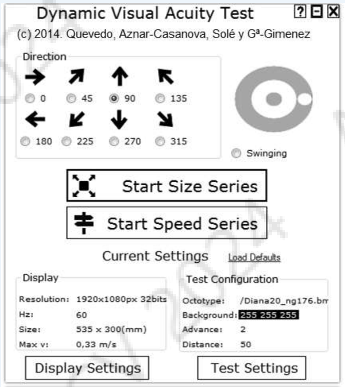
Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

Bernell's Rotator

The dynamic visual acuity values are recorded as a combination of **visual acuity** and **speed** in rpm.





Dynamic Visual Acuity Test

(c) 2014. Quevedo, Aznar-Casanova, Solé y Gª-Gimenez

Direction

○ 0 ○ 45 ● 90 ○ 135

○ 180 ○ 225 ○ 270 ○ 315

○ Swinging

Current Settings

Display

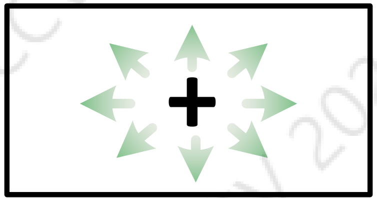
Resolution: 1920x1080px 32bits
Hz: 60
Size: 535 x 300(mm)
Max v: 0,33 m/s

Test Configuration

Octotype: /Diana20_ng176.bm
Background: 255 255 255
Advance: 2
Distance: 50

The DynVA is a computer software designed to assess DVA. The researcher can select the optotype to be presented in the two forms of the test: **(a)Size Series; (b) Speed Series**.

Human



Neuroscience:
simple symbol on high-contrast background

Different

Machine

Computer vision:
various targets in large-scale datasets

S. Hu, X. Zhao, Y. Wang, et al, "Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.

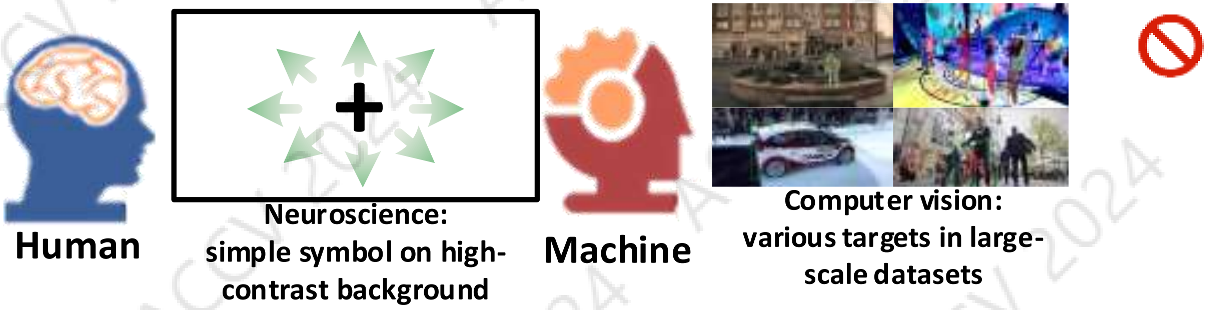
S. Hu, X. Zhao, and K. Huang, "Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.

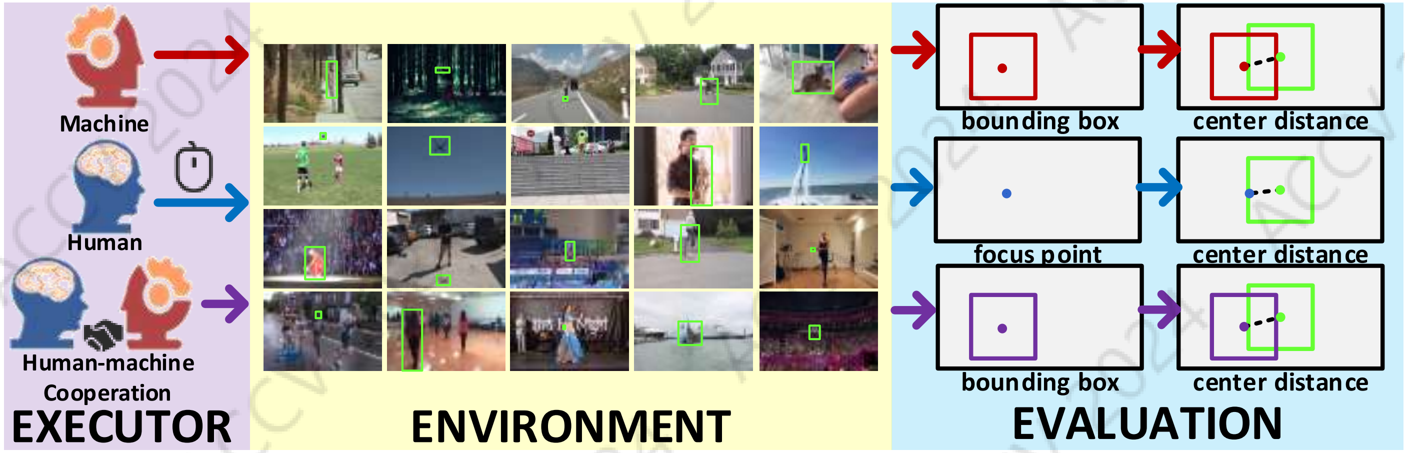
ACCV 2024

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Existing research**: Integrating human-machine evaluation into a unified framework for comparison and analysis is challenging due to **discrepancies across various research areas**.


Human
Neuroscience: simple symbol on high-contrast background
Machine
Computer vision: various targets in large-scale datasets


Machine
Human
Human-machine Cooperation
EXECUTOR
ENVIRONMENT
bounding box
center distance
focus point
center distance
bounding box
center distance
EVALUATION

➢ **Keypoint:** Designing a universal framework for evaluating dynamic visual abilities in humans and machines.

S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

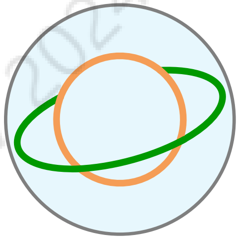
S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

ACCV 2024

Trend 4. More Intelligent Evaluation

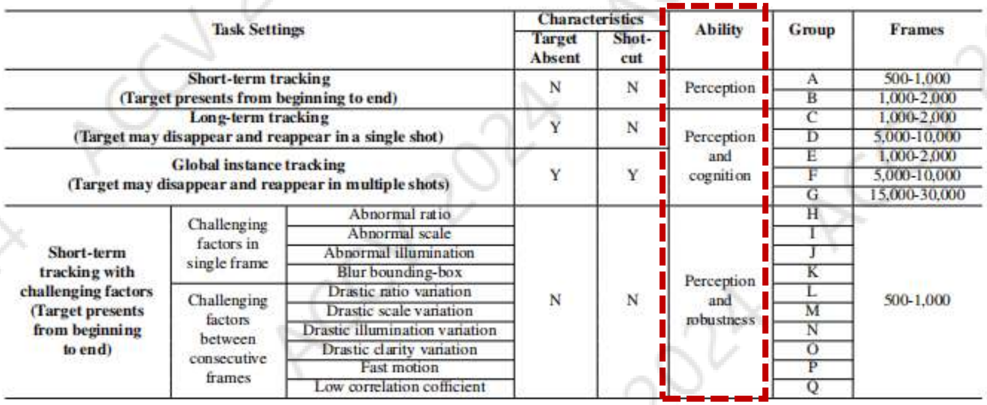
- **How to evaluate dynamic visual intelligence?**

  - **Environment:**
    - Provide a thorough evaluation environment of the **perceptual, cognitive**, and **robust** tracking abilities of humans and machines.
      - ☐ **87 sequences, 17 themes, 245k frames**

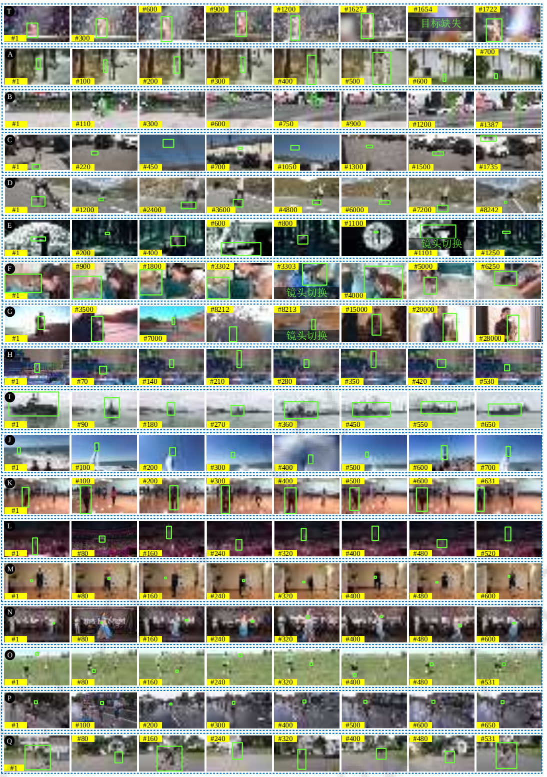


quickly build a task space based on research goals

Table 1: Information on environment settings.



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	



ACCV 2024

S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Executors:**

- 20 representative algorithms (different architecture)
- 15 human subjects were selected to participate in the visual tracking tasks, and their **behavior** was recorded (with a **self-developed program** by python)

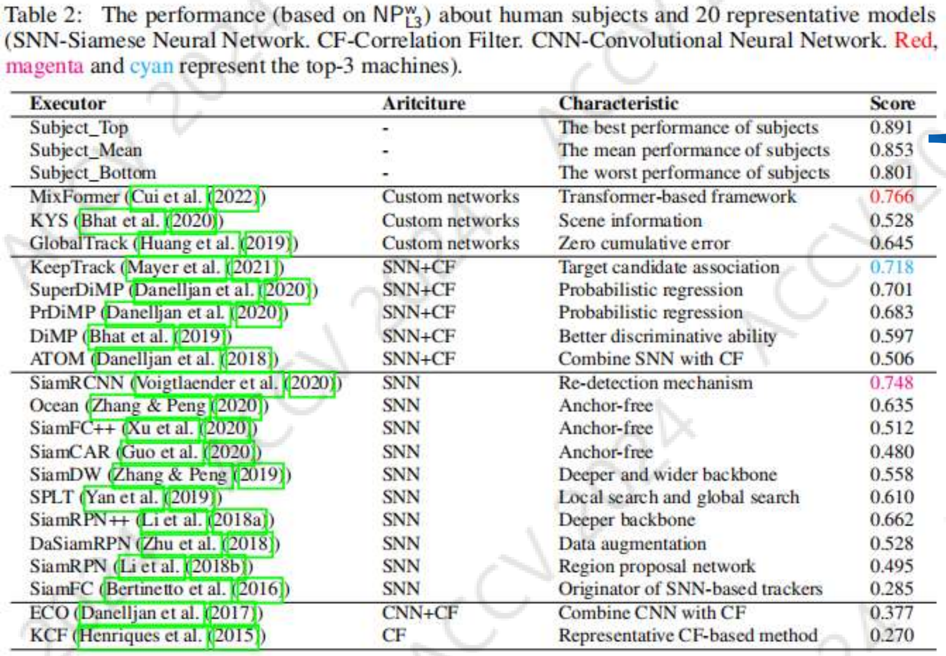
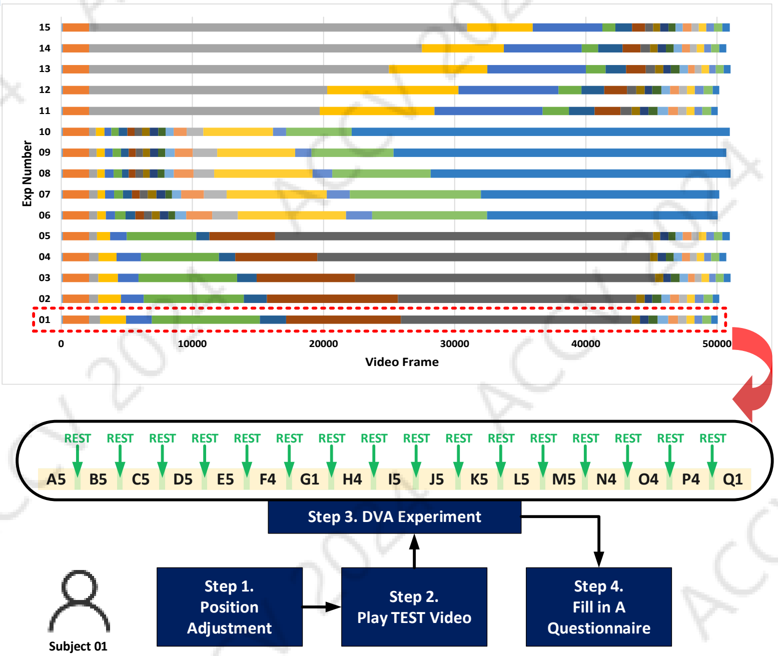


Table 2: The performance (based on NP^{w}_{L3}) about human subjects and 20 representative models (SNN-Siamese Neural Network. CF-Correlation Filter. CNN-Convolutional Neural Network. Red, magenta and cyan represent the top-3 machines).

Executor	Aritciture	Characteristic	Score
Subject_Top	-	The best performance of subjects	0.891
Subject_Mean	-	The mean performance of subjects	0.853
Subject_Bottom	-	The worst performance of subjects	0.801
MixFormer (Cui et al. (2022))	Custom networks	Transformer-based framework	0.766
KYS (Bhat et al. (2020))	Custom networks	Scene information	0.528
GlobalTrack (Huang et al. (2019))	Custom networks	Zero cumulative error	0.645
KeepTrack (Mayer et al. (2021))	SNN+CF	Target candidate association	0.718
SuperDiMP (Danelljan et al. (2020))	SNN+CF	Probabilistic regression	0.701
PrDiMP (Danelljan et al. (2020))	SNN+CF	Probabilistic regression	0.683
DiMP (Bhat et al. (2019))	SNN+CF	Better discriminative ability	0.597
ATOM (Danelljan et al. (2018))	SNN+CF	Combine SNN with CF	0.506
SiamRCNN (Voigtlaender et al. (2020))	SNN	Re-detection mechanism	0.748
Ocean (Zhang & Peng (2020))	SNN	Anchor-free	0.635
SiamFC++ (Xu et al. (2020))	SNN	Anchor-free	0.512
SiamCAR (Guo et al. (2020))	SNN	Anchor-free	0.480
SiamDW (Zhang & Peng (2019))	SNN	Deeper and wider backbone	0.558
SPLT (Yan et al. (2019))	SNN	Local search and global search	0.610
SiamRPN++ (Li et al. (2018a))	SNN	Deeper backbone	0.662
DaSiamRPN (Zhu et al. (2018))	SNN	Data augmentation	0.528
SiamRPN (Li et al. (2018b))	SNN	Region proposal network	0.495
SiamFC (Bertinetto et al. (2016))	SNN	Originator of SNN-based trackers	0.285
ECO (Danelljan et al. (2017))	CNN+CF	Combine CNN with CF	0.377
KCF (Henriques et al. (2015))	CF	Representative CF-based method	0.270



Step 3. DVA Experiment

Step 1. Position Adjustment

Step 2. Play TEST Video

Step 4. Fill in A Questionnaire

S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

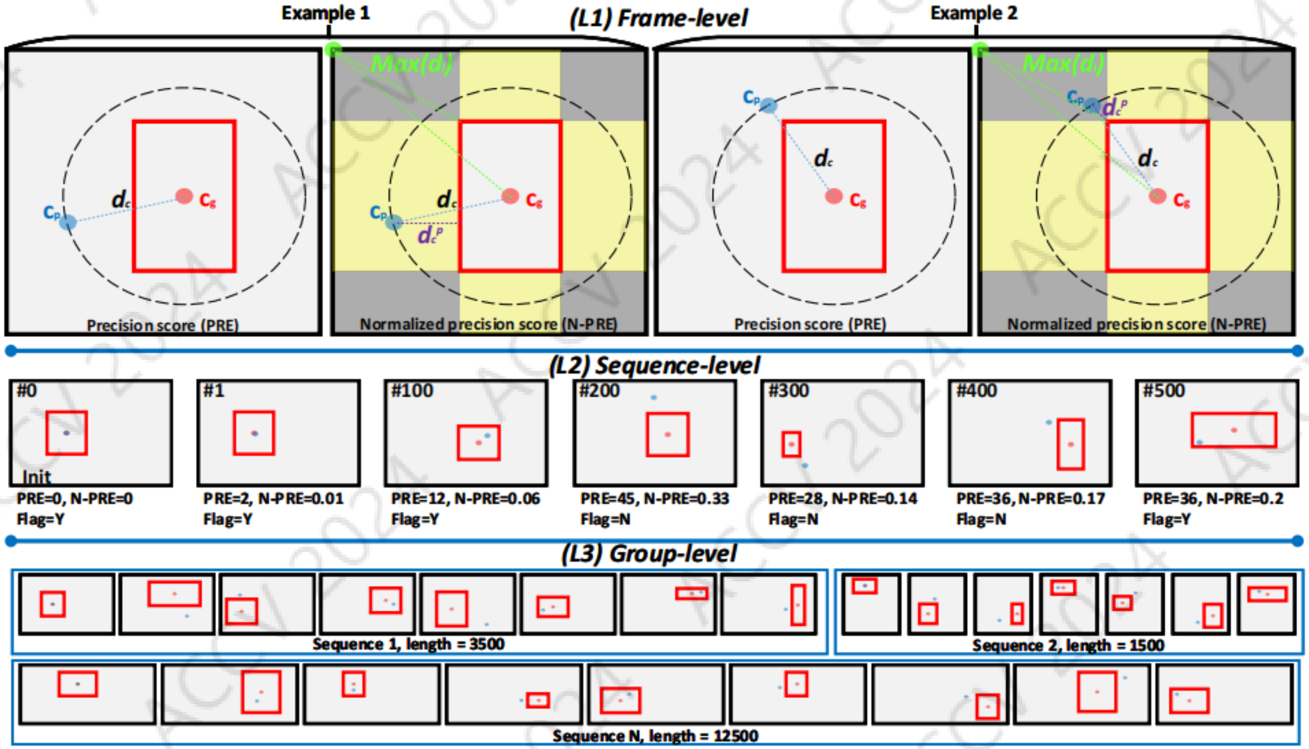
S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Evaluation:**

- Provide **universal multi-granularity evaluation indicators** (frame → sequence → group) for humans and machines tailored to task characteristics.


Example 1
(L1) Frame-level
Example 2
Max(d)
c_p
d_c
c_g
d_c^p
Precision score (PRE)
Normalized precision score (N-PRE)
Precision score (PRE)
Normalized precision score (N-PRE)
(L2) Sequence-level
#0
Init
PRE=0, N-PRE=0
Flag=Y
#1
PRE=2, N-PRE=0.01
Flag=Y
#100
PRE=12, N-PRE=0.06
Flag=Y
#200
PRE=45, N-PRE=0.33
Flag=N
#300
PRE=28, N-PRE=0.14
Flag=N
#400
PRE=36, N-PRE=0.17
Flag=N
#500
PRE=36, N-PRE=0.2
Flag=Y
(L3) Group-level
Sequence 1, length = 3500
Sequence 2, length = 1500
Sequence N, length = 12500

S. Hu, X. Zhao, Y. Wang, et al, "Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.

S. Hu, X. Zhao, and K. Huang, "Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.

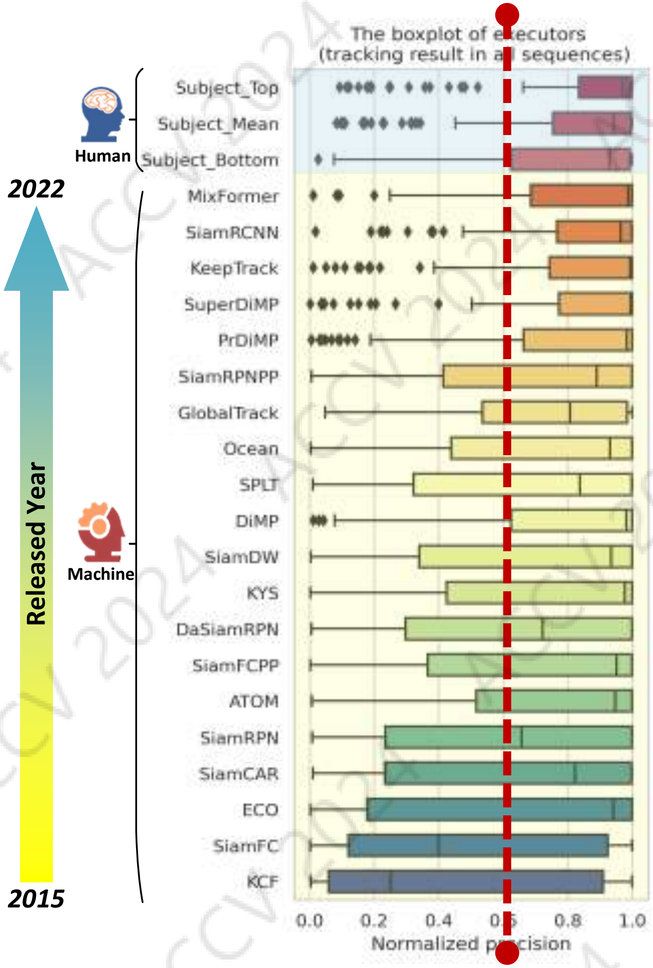
ACCV 2024

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Comprehensive comparison of human-machine dynamic vision capabilities:**

  - Human dynamic vision ability is **overall better** than most algorithms.
  - The current SOTA algorithm is close to the lower bound of human capabilities, and **the gap between the two is narrowing**.

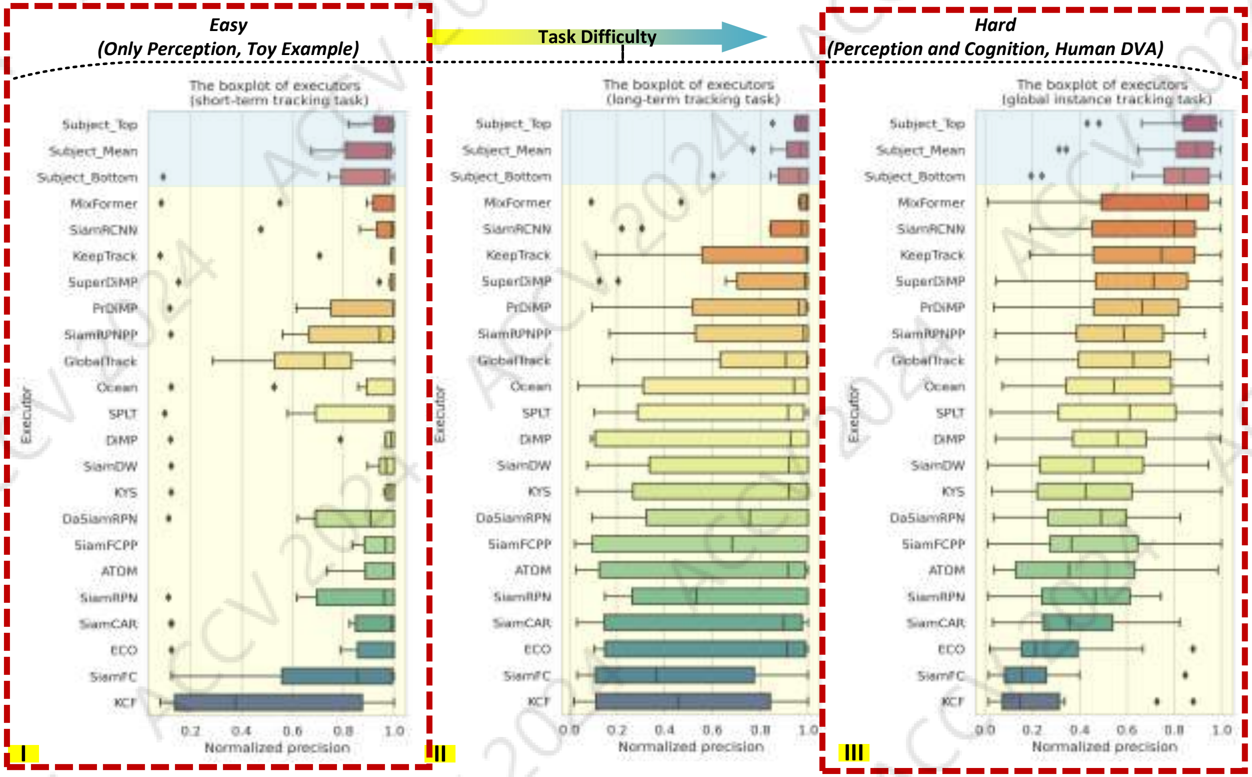


S. Hu, X. Zhao, Y. Wang, et al, "Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.

S. Hu, X. Zhao, and K. Huang, "Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**
  - ➢ **Comprehensive comparison of human-machine dynamic vision capabilities:**
    - Algorithms are **similar in perception** to humans.
    - There is still **a certain gap in cognitive abilities** between algorithms and humans.



S. Hu, X. Zhao, Y. Wang, et al, "Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.

S. Hu, X. Zhao, and K. Huang, "Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.

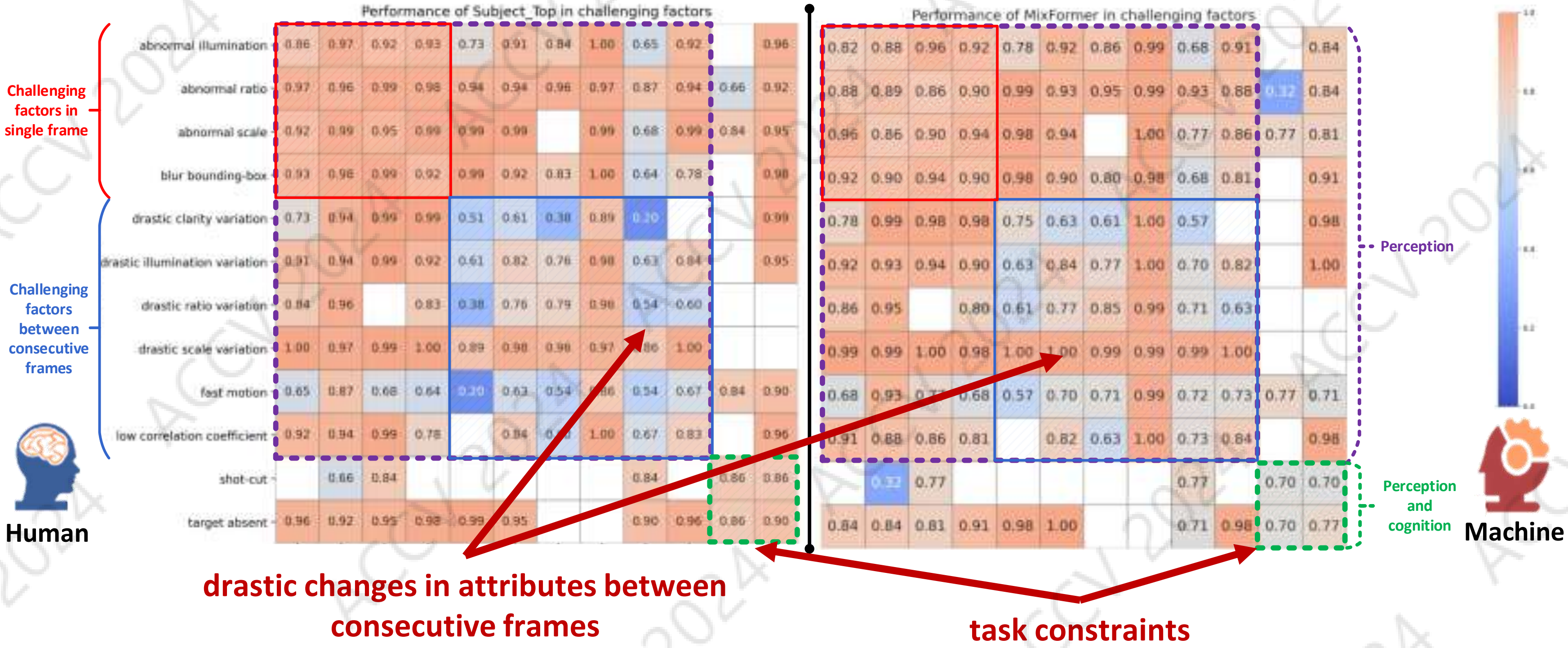
ACCV 2024

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Comprehensive comparison of human-machine dynamic vision capabilities:**

  - **Task constraints (such as camera switching)** have a greater impact on the machine.
  - **Drastic changes in attributes between consecutive frames** (such as fast motion) are challenging for both humans and machines.



abnormal illumination	0.86	0.97	0.92	0.93	0.73	0.91	0.84	1.00	0.65	0.92		0.96
abnormal ratio	0.97	0.96	0.99	0.98	0.94	0.94	0.96	0.97	0.87	0.94	0.66	0.92
abnormal scale	0.92	0.99	0.95	0.99	0.99	0.99		0.99	0.68	0.99	0.84	0.95
blur bounding-box	0.93	0.98	0.99	0.92	0.99	0.92	0.83	1.00	0.64	0.78		0.98
drastic clarity variation	0.73	0.94	0.99	0.99	0.51	0.61	0.38	0.89	0.20			0.99
drastic illumination variation	0.91	0.94	0.99	0.92	0.61	0.82	0.76	0.98	0.63	0.84		0.95
drastic ratio variation	0.84	0.96		0.83	0.38	0.76	0.79	0.98	0.54	0.60		
drastic scale variation	1.00	0.97	0.99	1.00	0.89	0.98	0.98	0.97	0.86	1.00		
fast motion	0.65	0.87	0.68	0.64	0.20	0.63	0.54	0.86	0.54	0.67	0.84	0.90
low correlation coefficient	0.92	0.94	0.99	0.78		0.84	[illegible]	1.00	0.67	0.83		0.96
shot-cut		0.66	0.84						0.84		0.86	0.86
target absent	0.96	0.92	0.95	0.98	0.99	0.95			0.90	0.96	0.86	0.90

0.82	0.88	0.96	0.92	0.78	0.92	0.86	0.99	0.68	0.91		0.84
0.88	0.89	0.86	0.90	0.99	0.93	0.95	0.99	0.93	0.88	0.32	0.84
0.96	0.86	0.90	0.94	0.98	0.94		1.00	0.77	0.86	0.77	0.81
0.92	0.90	0.94	0.90	0.98	0.90	0.80	0.98	0.68	0.81		0.91
0.78	0.99	0.98	0.98	0.75	0.63	0.61	1.00	0.57			0.98
0.92	0.93	0.94	0.90	0.63	0.84	0.77	1.00	0.70	0.82		1.00
0.86	0.95		0.80	0.61	0.77	0.85	0.99	0.71	0.63		
0.99	0.99	1.00	0.98	1.00	1.00	0.99	0.99	0.99	1.00		
0.68	0.93	[illegible]	0.68	0.57	0.70	0.71	0.99	0.72	0.73	0.77	0.71
0.91	0.88	0.86	0.81		0.82	0.63	1.00	0.73	0.84		0.98
	0.32	0.77						0.77		0.70	0.70
0.84	0.84	0.81	0.91	0.98	1.00			0.71	0.98	0.70	0.77

drastic changes in attributes between consecutive frames

task constraints

S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

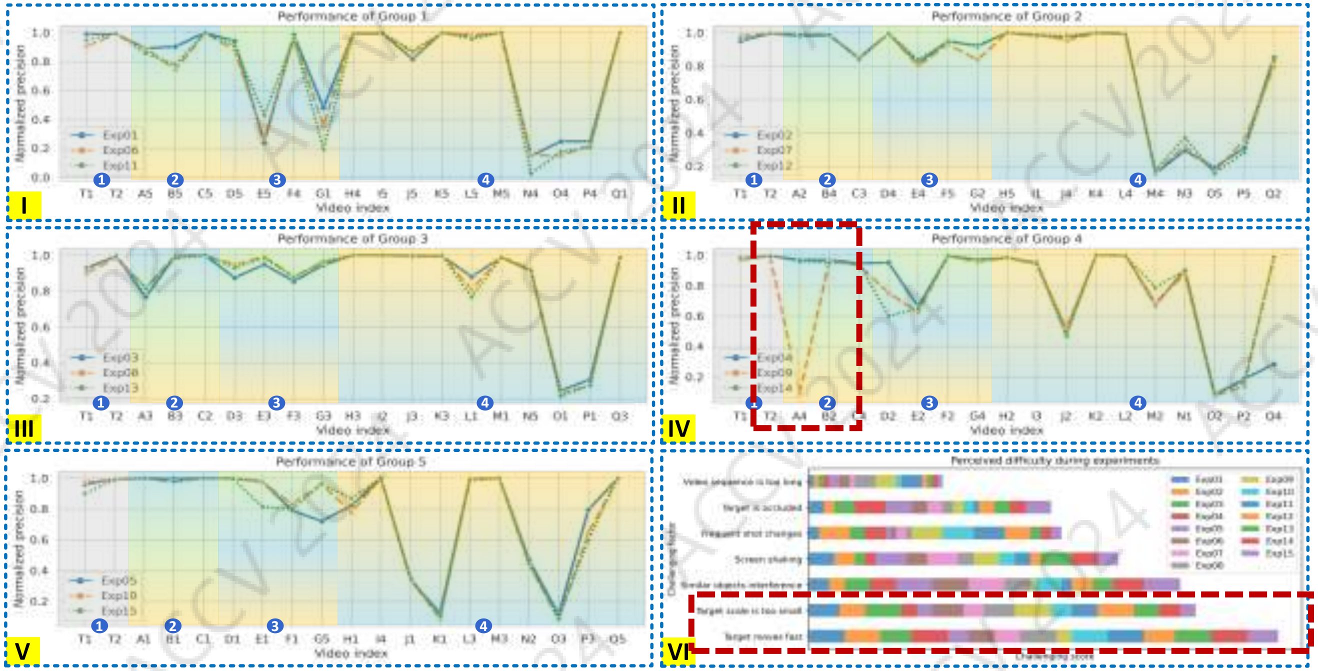
S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **Human Subject Performance Analysis:**

  - Human subjects also **make careless mistakes**.
  - The questionnaire showed that most subjects found it **difficult to track fast-moving targets and small targets.**



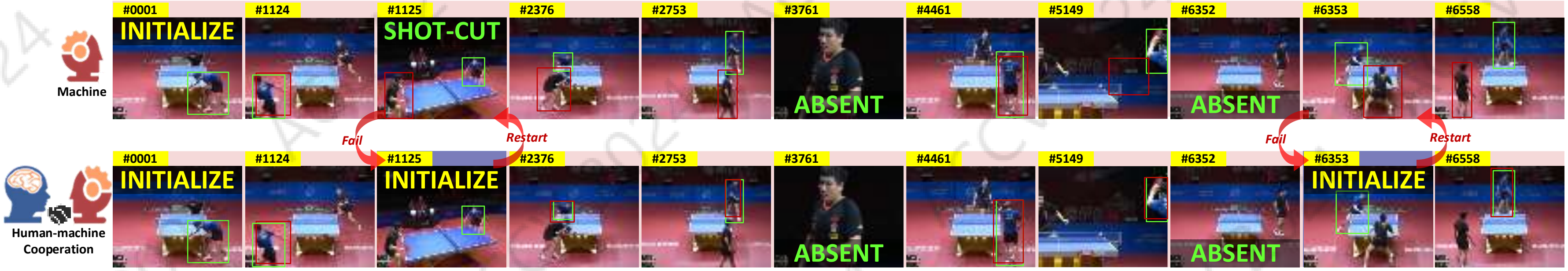
S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

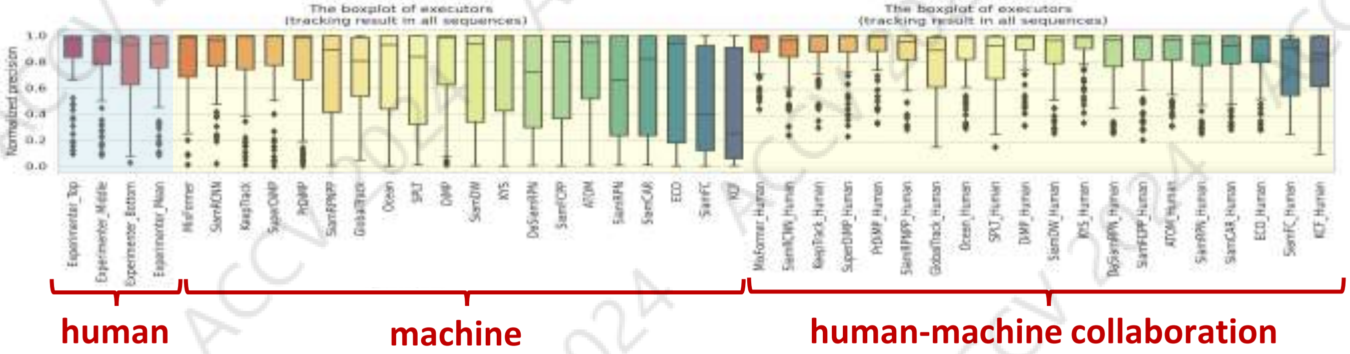
Trend 4. More Intelligent Evaluation

- **How to evaluate dynamic visual intelligence?**

➢ **A simple human-machine collaboration experiment:**


Machine
#0001
INITIALIZE
#1124
#1125
SHOT-CUT
#2376
#2753
#3761
ABSENT
#4461
#5149
#6352
ABSENT
#6353
#6558
Fail
Restart
Fail
Restart
Human-machine Cooperation
#0001
INITIALIZE
#1124
#1125
INITIALIZE
#2376
#2753
#3761
ABSENT
#4461
#5149
#6352
ABSENT
#6353
INITIALIZE
#6558

- A simple human-machine collaboration experiment shows that dynamic visual capabilities: **machine < human < human-machine collaboration**.
- In dynamic vision tasks, humans and machines each have their own strengths and **have the possibility of collaboration**.



ACCV 2024

S. Hu, X. Zhao, Y. Wang, et al, *"Nearing or surpassing: Overall evaluation of human-machine dynamic vision ability," Preprint, 2023.*

S. Hu, X. Zhao, and K. Huang, *"Visual intelligence evaluation techniques for single object tracking: A survey (单目标跟踪中的视觉智能评估技术综述)," Journal of Images and Graphics (《中国图象图形学报》, Top Chinese Journal), 2023.*

Future Work

Evaluation **Algorithm**

Decoupling visual capabilities:

- The visual object tracking task involves the coupling of multiple capabilities such as **observation, memory, and reasoning.**
- Therefore, the task can be further **decomposed**, and the intelligence of the algorithm can be more comprehensively analyzed and evaluated through a **fine-grained evaluation scheme**.

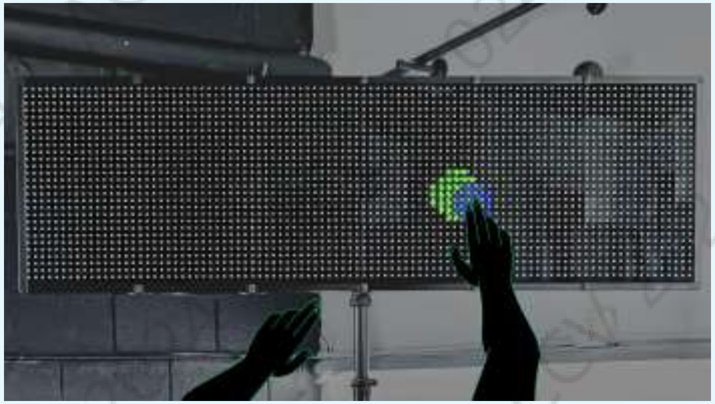
ACCV 2024

Future Work

Evaluation **Algorithm**

Optimize the measurement method:

- Use a high-precision eye tracker and set up a rigorous eye movement experiment environment, or design a new human visual tracking ability measurement solution (such as conducting experiments based on a **mouse** or **touch screen**).



ACCV 2024

Future Work

Evaluation ⚖ **Algorithm**

Exploring the characteristics of the subjects:

- Studies have shown that factors such as the subjects' **physiological characteristics, cognitive state, and personal traits** all have a certain impact on dynamic visual ability.
- How to select task objects based on the characteristics of the subjects to ensure that **the subject group is representative** is worthy of further analysis by researchers.

ACCV 2024

Future Work

Evaluation **Algorithm**

Visual modality target tracking algorithm:

- It can explore **a better mechanism for utilizing dynamic visual information** and strike a balance between the effective utilization of accumulated errors and temporal dependencies.

ACCV 2024

Future Work

Evaluation

Algorithm

Multimodal target tracking algorithm:

- Mature basic models in the fields of natural language processing and static vision can be introduced to improve the limitations of the algorithm in **long text processing and multimodal information alignment**.

ACCV 2024

Future Work

Evaluation

Algorithm

Expanding the human-machine collaboration mechanism:

- The human-machine collaboration mechanism can be further expanded to provide support for downstream tasks and practical applications. For example, **multiple rounds of human-machine interaction** experiments can be conducted during a single tracking process to observe whether the machine model's understanding of human intentions changes during the tracking process.

ACCV 2024

ACCV

北京科技大学
University of Science and Technology Beijing

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Thanks for listening!

2024.12.09 in Hanoi, Vietnam

Prof. Xin Zhao (赵鑫)

- Professor in University of Science and Technology Beijing (USTB)
- https://www.xinzhaoai.com/
- xinzhao@ustb.edu.cn

Dr. Shiyu Hu (胡世宇)

- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

Scan to download this tutorial PPT

ACCV 2024