Methodological Insights Towards Leveraging Performance in Video Object Tracking and Detection

Divyaprabha[1], Dr. M.Z Kurian[2]
Research Scholar[1], Professor[2]
Dept. of Electronics & Communication Engg, Sri Siddhartha Institute of Technology, Tumkur, India[1, 2]

***Abstract*—Video Object Detection and Tracking (VODT), one of its integral operations of surveillance system in present time, mechanizes a way to identify and track the target object autonomously and seamlessly within its visual field. However, the challenges associated with video feeding are immensely high, and the scene context is out of human control, posing an impediment to a successful model of VODT. The presented work has discussed about effectiveness of existing VODT approaches considering its identified taxonomies viz. satellite based, remote sensing-based, unmanned-based, Real-time Tracking based, behavioral analysis and event detection based, integration of multiple data sources, and privacy and ethics. Further, research trend associated with cumulative publications and evolving methods to realize the frequently used methodologies in VODT. Further, the results of review showcase that there is prominent research gap of manifold attributes that demands to be addressed for improving performance of VODT.**

Keywords*—Object detection; object tracking; video; visual field; surveillance system; video feed

I. Introduction

Object detection and tracking are essential operations that any surveillance system demands [1]. Irrespective of archives of research models presented to date, there is still an open concern associated with object detection and tracking [2]-[5]. One of the significant shortcomings of cumulative research work is the lack of any model which can guarantee maximized performance, higher accuracy and dominantly robust [6]-[8]. The prime challenges are encountered in this domain mainly because of the condition stated towards tracking and detecting an object for a given scenario. The complexity of such implementation depends upon the use cases. In the case of objects with fewer visual features, it is subjectively easier to detect all the features associated with that visual characteristic. However, complexity arises for the object of dynamic type, where extraction of features is quite a complex process in the presence of challenging background and foreground situations. Predominant challenges surface regarding Video Object Detection and Tracking (VODT), especially for moving objects in general. However, there could be multiple mobility situations and statics of either foreground or background [9]. Various other factors that impose challenges in VODT are occlusion (full/partial), fluctuation in illumination condition, deformation of the target, and variability in the pose of the target object. Various standard approaches of detection consist of object detection based on i) features (color, shape) [10], ii) template (deformable / fixed) [11], and iii) motion (global energy, statistical test, thresholding) [12]. On the other, the approaches of tracking are carried out using motion information mainly (region, boundary) [13]. The conventional mechanism of VODT emphasizes two essential attributes, i.e., information related to the motion of the target object and visual features (e.g., shape, texture, color, etc.). Owing to the variable nature of such attributes, it is always better to perform the modeling of VODT by integrating temporal features and statistical models with visual features [14]. Normally, the process of VODT consists of obtaining regions based on visual features using frame segmentation followed by combining all regions characterized by equivalent motion vectors. The majority of the existing approaches of VODT are witnessed with the adoption of multiple forms of approach as well as a combination of approaches. The area of implementation is so scattered that there is a lack of any uniform approach towards VODT with a consistent track of performance. It is essential to offer a comprehensive insight into various study models towards VODT to identify the possibilities of future research work by reviewing its strength and weakness. Therefore, the proposed study discusses existing approaches to offer insight into the effectiveness of existing VODT. The contributions made in the proposed paper are:

- Commercial usage of VODT is discussed to understand global deployment in the practical world,
- Existing approaches are classified concerning some potential use cases of deployment to understand the effectiveness in deployment scenarios,
- Discussion of research trends is carried out to offer insight into identifying frequently used techniques,
- A crisp discussion of the research gap associated with existing VODT techniques.

The organization of this paper is as follows: Section II discusses insight into the commercial application of VODT, followed by an elaborated discussion with the classification of VODT approaches in Section III. A discussion of the most frequently adopted video dataset is carried out in Section IV. At the same time, the research trend is discussed in Section V. Highlights of the results discussion is carried out in Section VI, while conclusive remarks are stated in Section VII.

II. Commercial Usage of VODT

The commercial usage of ODT mainly performs localization, identification, tracking, counting, and recognizing the anomalies of an object of specific form present within the captured image frame of the Video. The ranges of such objects

are quite wide enough. The devices that perform such tasks consist of recognition and classification modules. In general, ODT from the video scene is carried out by extracting the object from the background image, presenting an anticipated equivalent object class proposition. Finally, a bounding box is constructed to encapsulate the object. This section discusses the different use cases of the application of ODT.

A. Smart City Usage

With the advent of the Internet-of-Things (IoT), the environment of a smart city is now equipped with multi-functional sensors and sophisticated devices for monitoring. Various ODT applications find a suitably higher scope in smart cities, viz.

- Monitoring of In-Cabin Space: Modern technologies are now used for monitoring the environment inside a closed space like a home or vehicle to understand people's behavior. This is done using computer vision, eye tracking, pose estimation, etc. It is specifically helpful in detecting drivers' drowsiness to make the ride safer [15].
- Occupancy of Parking Site: The mechanism of VODT in computer vision is used for classifying vacant and occupied parking areas. The information can be transmitted in real-time directly to the driver directing them towards the target open end parking space [16].
- Counting People: This application tracks visitors' counts to plan security in public areas. They are usually deployed in transit areas, which can capture data on incoming and outgoing people for security and capacity management [17].
- Monitoring Traffic: Various Road and traffic conditions can be monitored via VODT. Multiple zones in the city can be monitored to identify congestion or violation in driving or accidents. It is also used for identifying the vehicle's registration details by capturing the license plate image [18].
- Autonomous Driving: Unmanned driving vehicle is the next vehicle level with the advancement of computer vision technology. Such application uses VODT to identify traffic signs, markers on the lane, vehicles, pedestrians, etc. [19].

B. Industrial Usage

ODT is also used for various industrial applications on a large and small scale. The following are the different uses of VODT:

- Enhancing Productivity: Multiple activities of the workers in different locations, e.g., construction sites, production facilities, and warehouses, can be monitored using VODT. The worker's activities can be monitored while retaining private information [20].
- Detection of Defect & Anomaly: Different forms of industrial products can be assessed using computer vision to identify the defects or sub-standard quality in finished products. Various processes associated with quality control, production lines, and workstations can be monitored using VODT [21].
- Product Assembly: Adoption of ODT can be used for ensuring the selection of appropriate components over the assembly lines. Different forms of automated production systems and robotics are facilitated with various intelligent feeds from the ODT system [22].
- Detection of PPE: ODT can also be used to ensure worker safety. This is done by evaluating if the employees have put on a specific form of Personal Protective Equipment (PPE) assigned by the safety standards. The system can distinguish between employees with and without protection gears [23].

C. Retail Usage

The usage of computer vision plays an important role in the retail industry with multiple purposes. The retail sector has manifold concerns about reviewing the user's response, behavior, sales processes, etc. Various retail-based applications are:

- Customer Experience: ODT can extract the customer's actions during store visits. It can assist in understanding the possibilities of assistance that any visiting customer may require [24].
- Analysis of Foot Traffic: This is nearly similar to object counting applications under smart city usage. This application is used for counting the incoming and outgoing visitors in the store to understand the peak traffic of visitors. The manager can directly analyze this traffic information, assisting in product placement or promoting the product at a specific location [25].
- Inventory Management: This application is used for evaluating the availability of products on specific shelves or warehouses. In case of a drop in product availability, the ODT-based system can forward the notification to the inventory manager to restock. Artificial intelligence effectively controls inventory management systems [26].
- Contactless Checkout: Contactless kiosk is used for a contactless payment system, an effective solution for crowd and queue management inside the store [27].
- Video Analytics: ODT can be used for capturing the feed inside retail to evaluate customers' shopping behavior. Potential analytical information can be obtained from such form of Video feeds powered by artificial intelligence. It can also assist in faster and more effective service delivery within the store with less customer wait time [28].

Therefore, all the taxonomies mentioned above of application for VODT are some of the prime focuses on commercial and research interests. The next section discusses the current approaches of VODT.

III. Current Related Work of VODT

At present, there are various categories of approaches as well as methodologies that have been undertaken towards investigating VODT problems. However, categorization concerning methodologies is quite a tedious process as it is noted that sometimes the same methodology is adopted to solve two different use cases of problems. Hence, this part of the discussion is based on various use cases of problems being investigated and highlights individual methodologies used to address the problem.

A. Satellite based VODT Approaches

This mechanism calls for capturing the satellite's video feed and applying algorithms to identify the objects from the feeds. It is one of the most evolving and challenging investigation trends in VODT, irrespective of progressive research [29]. Most of the existing research approach targets solving the identification and tracking of smaller objects from satellite video datasets. The prime challenge is to identify the foreground objects, which are smaller and have low contrast. The work carried out by Chen et al. [30] has developed a scheme for enhancing the tracking performance of an object of smaller sizes captured from satellite-annotated images. Hu et al. [31] have constructed a network based on a regression model integrated with gradient descent and convolution layers. The tracker is designed from the background context using a regression network. A deep neural network is used to train motion and appearance features. Shi et al. [32] have implemented a technique for detecting and tracking mobile aircraft. The investigation carried out by Wu et al. [33] constructed an enhanced filter with kernel correlation to track the smaller-sized object. According to this experiment, the mean peak response is integrated with the mean energy of peak correlation associated with the response map to mitigate occluded objects.

Further, an adaptive Kalman filter is used to improve the tracking performance. Xuan et al. also uses a similar methodology line [34] where the motion trajectory is integrated with the Kalman filter to track the smaller object. The study outcome is claimed of 95% accuracy in tracking performance. The work by Zhang et al. [35] has integrated features of optical flow with a Histogram of oriented Gradient (HoG) with a target towards enhancing the target representation. The boundary effects are mitigated by integrating the inertial mechanism and Kalman filter, while the interference attenuation is accomplished using the disruptor-aware process. Zhou et al. [36] performed a unique study considering satellite images where a pyramid network of selected features is presented to address the problem of computing gradient inconsistencies. The study has also introduced a contrastive learning mechanism to represent an object robustly. The work by Zhu et al. [37] used Siamese deep network designed to improve the smaller object representation from satellite videos. This model aims to extract features from the search and template branch in the Siamese network while the target position is determined from the search branch. The model's outcome is proven to offer a satisfactory accuracy evaluated over multiple performance parameters.

B. Remote Sensing-based VODT Approaches

This mechanism identifies and evaluates the physical characteristic associated with a region by computing the radiation emitted from a specific distance. Usually, such monitoring is carried out from aircraft or satellites using specific feed-capturing devices. The present research is being carried out towards remote sensing-based VODT using a conventional mechanism [38] that consists of simplified object detection and tracking. Detection is facilitated by various techniques, e.g., background subtraction process, optical flow mechanism, and method of computing frame difference. The work carried out by Lei and Guo [39] implemented a technique that can detect and track multiple objects considering the Gaussian mixture model for extracting road networks using deep learning.

Further, a neighborhood search mechanism is implemented for tracking connected with the data association method. The neural network adoption is witnessed in Lin's [40] work, where tracking is carried out for multiple targets. The implementation is carried out considering the integration of real-time tracking of the target using deep learning, combined geometric features, and a dual neural network. The study model also constructs an optimization mechanism using the least squares method, where the constrained residual terms are developed over the pixel plane. A unit for inertial measurement is developed. Finally, a mathematical model is constructed for multitarget tracking. The mechanism presented by Ma [41] discusses enhancing the probability of identifying objects with poor signal quality and minimizing the number of outliers. Another unique study has been presented by Tochon et al. [42] on chemical gas plume tracking. Considering hyperspectral Video sequences, the author has used object detection sequentially, considering temporal, spatial, and spectral information. The work carried out by Uzkent et al. [43] [44], where a fusion mechanism of kinematic likelihood is estimated for analyzing hyperspectral information. The idea is to detect and track mobile objects captured from aerial Video over a short time window.

Further, the author has also used deep features for improving the tracking performance using a convolution neural network and kernelized correlation filter. However, the study model cannot process an isometric view of the scene. This problem is reported to be addressed in the work of Wei et al. [45], where a three-dimensional view is considered with total variation towards tracking a moving object.

C. Unmanned VODT Approaches

Various existing VODT use cases have considered the Video captured from Unmanned Aerial Vehicle (UAV) system. The work carried out by Cintas et al. [46] has constructed a model that can perform tracking of UAV using vision factor from another UAV using a deep learning approach and kernelized correlation filter. Deng et al. [47] used a regularization method using spatial and temporal attributes and constructing an enhanced filter for discriminative correlation. The technique also uses a joint optimization mechanism to improve target representation and reduce distractors. A unique method is implemented by Ding et al. [48], where object detection for UAVs is carried out by blockchain and hash-based approach. According to this implementation, a hash-

based network is constructed for extracting the hash representation of an object where further recovery of tracking interrupts and feature fusion is carried out. The study model also integrates motion features with deep hash features to address the problem of object occlusion. Detection of a small object is carried out by Liang et al. [49] using a fusion of features and Detection of a single shot based on scaling operation. A feature pyramid is constructed using average pooling operation. The fusion of the feature module and the deconvolution module is used for generating the feature pyramid. Lin et al. [50] have designed a correlation filter based on blocks' bidirectional incongruity towards tracking objects. Buhler et al. [51] have carried out a different form of work that doesn't work on video object detection or tracking algorithm but offers a robust platform to perform such tasks. The authors have used remote sensing images of snowflakes to classify snow types' normal to critical states. Shan et al. [52] have developed a ship-tracking mechanism considering maritime data over multiple videos with manual annotation. Xue et al. [53] have developed a semantic-based framework for object tracking to improve the performance of correlation filters of discriminative nature. The study model generates the specific region of interest followed by filtering where the semantic coefficient prediction is carried out. The study model also contributes towards reducing parametric redundancy by sharing the object's semantic segmentation information and network layer. The adoption of a correlation filter is also witnessed in the work of Ye et al. [54]. The study model applies a concept of multi-regularization toward constructing a correlation filter to facilitate the tracking and localization of an object. The work carried out by Yu et al. [55] has developed a region of interest while object detection is carried out from radar and photoelectric cell.

D. Miscellaneous Approaches

Various methodologies are being introduced, considering different forms of use cases. The work carried out by Banerjee et al. [56] has developed a hardware design of architecture for the acquisition of Video. The study model uses dynamic programming towards the reduction of normalized area for the allocation of resources. Further, the Kalman filter is used for performing tracks. A simplified mechanism of object detection and tracking is carried out by Chen et al. [57], where a region of interest in a closed loop is developed to magnify the field of view, considering spatial and temporal attributes of the feed. Cheng et al. [58] have developed a framework to track commodities where the edge computation caters to resource dependencies. The model can also track multiple videos while significantly controlling computation overhead.

Further, the Markov model is designed for compensating the missing data. A similar form of edge computing investigation is also carried out by Gu et al. [59], where a Siamese convolution network is constructed to carry out object tracking. A collaborative architecture of cloud and edge computation is designed for this purpose. The adoption of the Siamese network is also reported in work carried out by Lee et al. [60] towards object tracking. According to this model, a bounding box encapsulates the object, followed by tracking using a Siamese network which minimizes the computational complexity. Object detection is carried out towards each classified object in Video, followed by performing tracking using sequential frames. All the obtained retargeted frames are obtained by rearranging all frames.

A discreet and unique model of object tracking is designed by Liu et al. [61], which addresses the problem of the extreme degree of occlusion. It also assists in the classification of similar forms of objects. The study model uses a depth tracking mechanism and presents a matching strategy of two look-alike objects using semantics using indoor scene video. The study by Marshall et al. [62] developed a three-dimensional tracking mechanism of an object of radiological or nuclear type. The study uses a Kalman filter to perform tracking using multiple cameras of a specific form. Mostafa et al. [63] have developed a computational model of multi-object tracking using the Kalman filter for tracking crowds. According to this model, a unique encoder model is designed to generate computationally efficient features.

Further, a linear transformation is carried out to retain maximum accuracy. Ramesh et al. [64] have presented a framework for long-term tracking over dynamic motion. Considering a usual tracking condition, the proposed scheme is implemented considering a moving camera where a local sliding window is formulated to confirm the higher reliability. The analysis considers quantitative analysis, rotation, and translation in a laboratory environment. The work carried out by Ren et al. [65] has discussed a computational model that can perform both tracks as well as counting from the crowd scene using a network flow approach. Sun et al. [66] have implemented a study model for tracking multiple objects. The presented study uses deep learning to optimize the representation of an object and their respective affinities occurring over every frame. An affinity-based deep learning network is designed to track objects present and disappearing from the frames. Eventually, the problem of the missing object is addressed in this work.

E. Robust Real-Time Object Tracking

The practical applications of object tracking demand its operation of real-time feed of video stream which is one the most challenging task. At present, there are certain research work which has been dedicated purely towards ensuring robust real-time object tracking. The work carried out by Zhang and Ren [67] have used Kalman Filter and Kernel Correlation Filter in order to carry out object detection. Further, missed detection is addressed by updating the tracking box as per the rate of change in scaling. Backstepping is used for catering the design of kinematic controller with respect to Lyapunov stability. Further, the work of Cao et al. [68] have addressed the issues associated with tracking drift and loss of object by using Siamese network with double template while the feature extraction is performed by enhanced MobileNet V2. Similar work is also carried out by Zhao et al. [69] where Siamese Network is used for estimation of optical flow based on feature pyramid approach. The movement attributes are further evaluated using mapping of pyramid correlation between two contiguous frames. The study also addresses the problems associated with ambient noise using channel and spatial attention. Another study of real-time object tracking is carried out by Du et al. [70] that addresses the problems associated with implementing Correlation Filter-based Trackers (CFT) on

practical grounds. In this mechanism, feature extraction is carried out using Histogram of oriented Gradient (HoG) followed by integration of scale adaptive, target re-detection, and discriminative appearance model. Although, there are various research work carried out towards object tracking, but only few of the recent work has been claimed towards real-time tracking. Another perspective is that these studies doesn't address overall agenda-based attributes that is demanded for real-time tracking e.g., initialization and re-detection, computational constraint, fast motion, motion blur, perspective and scale changes, and occlusion all together. Only few of the above-mentioned attributes have been choses in current works on real-time object tracking.

F. Behavioral Analysis and Event Detection

There are various unique studies being carried out towards object detection in the perspective of event detection and behavioral analysis. The work of Chakole et al. [71] has addressed the issues pertaining to anomaly detection for the use-case of crowd behavior using correlation-based optical flow. Object detection has also been studied with respect to recognition mechanism of human activity. A typical structure of recognition of human activity is designed by Vrskova et al. [72] using deep learning approach. The prime motive of this work is to address the challenge associated with the less availability of such dataset, where the authors have constructed a new dataset consisting of abnormal activities. Study associated with anomaly detection is carried out by Chang et al. [73], Yahaya et al. [74], and Yang et al. [75]. Such methodologies are meant to identify a discrete vector of motion, which is further subjected to varied strategies in order to perform behavioral analysis. Irrespective of varied approaches used for such form of behavioral analysis, they still have limiting attributes associated with their practical utilization on real-time perspective.

G. Integration of Multiple Data Sources

Majority of the conventional implementation scenario calls for implementing its detection and tracking algorithm considering single sources of data. However, there are certain research work where such objectives are accomplished using multiple sources of data. One such work has been reported by Rehman et al. [76] where visual and audio signals were used as input stream for detection of anomaly. According to this study, the model has integrated particle swarm optimization with optical flow for obtaining visual features. The acoustic features have been obtained from volume, rate of zero crossing, energy, spectral flux etc. Similar pattern of methodology is also created by Benegui and Ionescu [77] which carry out authentication using feeds from camera and motion sensors. Deep neural networks are used using Support Vector Machine for generating an embedding vector obtained from transformed signals. Another similar form of study approach is seen in work of Shin et al. [78] where heterogeneous sources of data e.g., temperature, elevation changes, pattern-specific behavior of human have been considered. The prime notion of this study is to improve the accuracy towards detection of anomaly associated with surveillance map. Majority of the above-mentioned studies have been carried out by integrating multiple and different sources of data in order to obtain an embedded vector which are further subjected towards analysis towards detection and tracking.

H. Privacy and Ethics

Privacy is one of the essential attributes to be protected while performing object detection. The work carried out by Dave et al. [79] has developed a scheme towards privacy preservation while performing recognition of human action. The authors have used supervised schemes towards removing the private details without any dependency of labelling. The work carried out by He et al. [80] have implemented a mechanism of object blurring as well as object swapping in order to preserve private information while performing object detection. Zhang et al. [81] have used federated learning scheme along with blockchain for detection visual object along with retention of private information. This scheme uses encryption as well as validation nodes in order to resist any form of privacy-related attacks. Further mechanism of object detection along with privacy preservation is carried out by Bai et al. [82] where Convolution Neural Network (CNN) has been utilized along with secret sharing scheme towards protecting private information captured from vehicular edge computing. The prime motive of such approaches is basically to adopted varied privacy preservation approaches ensuring that it won't hamper any parameters potentially responsible for detection of an object.

Hence, it can be seen that there are various ranges of methodologies being developed in current times on various perspective of object detection. It is eventual that all these approaches leverage the performance of object detection addressing issues on local levels and use case.

Apart from this, various works is carried out in conventional VODT with different techniques. One common fact observed in all the miscellaneous techniques is that tracking is more emphasized than preliminary detection operation. As tracking performance completely depends upon the logic configured for the detection module, the performance evaluation doesn't highlight this fact. Apart from this, all the usual and generalized VODT schemes discussed in this section can be used for multiple purposes and assessed over different datasets; however, their applicability towards functional over complex scenarios or challenging video environments is reportedly not investigated. Another significant observation is that most schemes utilize varied principles of framing up the dimension of an object to track the objects in the video stream. However, there is no report of any work towards the generalized object tracking model utilized in the video stream to ensure better analytical operation toward improving the tracking performance. Further, it is noted that there is different use-case specific implementation study where generalization is challenging to be achieved. Table I highlights the summarized version of the discussion in this section concerning methods used for addressing identified research problems, advantages, and limitations.

TABLE I. SUMMARY OF RELATED WORK IN VODT

Author	Problem	Methodology	Advantage	Limitation
Chen et al. [30]	Tracking of the smaller object	Historical model, Kalman filter	Higher Accuracy	Lacks scale adaptive process
Hu et al. [31]	Drift in tracking	Convolution Neural Network, Regression	Better tracking performance	Orientation and position encoding are not feasible
Shi et al. [32]	Tracking mobile aircraft	Shift invariant feature transform	The better predictive calculation for flight path	Not applicable for multiple object tracking
Wu et al. [33]	Tracking of the smaller object	Adaptive Kalman filter	Mitigates occlusion problem	The iterative process leads to a computational burden
Xuan et al. [34]	Tracking of the smaller object	Kalman filter, averaging motion trajectory	95% of accuracy, effective processing speed	Restricted to single object tracking
Zhang et al. [35]	Tracking of the smaller object	Kalman filter, HoG, inertial mechanism	Can perform multi-object tracking, effective classification	Doesn't overcome occlusion problem
Zhou et al. [36]	Gradient inconsistencies	Pyramid network with selected feature scale	Effective reduction of interface	Doesn't consider temporal factors while tracking
Zhu et al. [37]	Inferior distinguishability of objects	Siamese deep network	Offers real-time processing capabilities	It doesn't address the impact of padding and shallow network
Lei and Guo [39]	Detection and tracking	Gaussian Mixture Model, Neighborhood Search	A simplified approach supports multi-object Detection	No benchmarking
Lin [40]	Multitarget tracking	Deep learning, dual neural network	Retain superior tracking quality	Narrowed extensive analysis
Ma [41]	Addressing Outliers in Detection	Probability-based modeling to reduce outliers	Improved rate of Detection	Lower scope of analysis
Tochon et al. [42]	Tracking of chemical gas	Sequential object tracking using temporal, spatial, and spectral information	Reduced computational time	Specific to the composition of chemical gas
Uzkent et al. [43][44]	Tracking of vehicle	Fusion mechanism for kinematic likelihood, convolution neural network, and kernelized correlation filter.	Independent of manual allocation of the threshold for multiple objects	No scope for an isometric view of the scene.
Wei et al. [45]	Object tracking	Principal component analysis	Simplified mechanism of implementation	No benchmarking
Cintas et al. [46]	Tracking UAV from another UAV	Kernelized correlation filter, neural network	82.7% of accuracy, developed a new dataset	Demands more real-time testing to confirm the outcome applicability
Deng et al. [47]	Limited computational capability in UAV tracking	Regularization, correlation filter, joint optimization	Better tracking performance	Higher processing time
Ding et al. [48]	Multiple object tracking	Blockchain, deep hash	Mitigates object occlusion	Consumed higher resources
Liang et al. [49]	Detection of a small object	Spatial context analysis	Higher accuracy	Demands higher resources to map with the outcome
Lin et al. [50]	Improving correlation filter	Discriminative correlation filter	Support real-time tracking with low resources	Outcome specific to data
Buhler et al. [51]	Snow type classification	The reflective property of snow-based modeling	Effective classification strategy	Study model specific to the object
Shan et al. [52]	Ship tracking	Multiple maritime dataset evaluation	Comprehensive analytical approach	Study model specific to the object
Xue et al. [53]	Correlation tracking	Semantic segmentation	Minimize parametric redundancy	It doesn't address the occlusion problem effectively
Ye et al. [54]	Improving correlation filter performance	Multi-regularized filter development	Higher accuracy	Not analyzed on broader assessment environment
Yu et al. [55]	Object tracking	Prediction based on the region of interest, object extraction from the first frame	Reduced processing time	Tracking performance degrades over dynamic background
Banerjee et al. [56]	Acquisition of object	Viterbi, Kalman filter quadtree segmented Video, convolution neural network	Extensive test cases	Needs a broader extensive assessment environment
Chen et al. [57]	Object detection & tracking	Region on interest on a closed loop, dual resolution	Higher resolution	The computationally extensive mechanism for identifying a region of interest
Cheng et al. [58]	Commodity tracking	Edge computing, Markov model	Reduces computational burden	It needs more extensive analysis with recent models
Gu et al. [59]	Object tracking	Siamese Convolution Network, cloud, and edge environment	Reduced energy consumption, Higher accuracy	No extensive analysis
Lee et al. [60]	Object detection	Siamese Convolution network	Higher similarity score	Highly iterative scheme
Liu et al. [61]	Multiple indoor object tracking	Semantic strategy to match objects	Simplified and efficient tracking	It depends upon the illumination condition
Marshall et al. [62]	Object detection	Kalman filter, neural network, non-negative matrix factorization	Reduces anomaly count	Applicability is Specific to this use case only

Mostafa et al. [63]	Crowd tracking	Encoder, Kalman filter	Control over computation time	It depends upon the illumination condition
Ramesh et al. [64]	Object tracking	Bayesian Bootstrapping, sliding window detector	Faster Detection	Assessed over laboratory confined setting.
Ren et al. [65]	Object tracking, Detection, counting	Network flow programming	Supports Multiple operations	The optimization of the counting problem is not addressed
Sun et al. [66]	Tracking multiple objects	Affinity-based deep learning,	Effective analysis	Induces computational complexity over the long run
Zhang and Ren [67]	Real-time object tracking for robots	Kalman Filter, Kernel Correlation Filter	Higher stability of target tracking	Doesn't assess model uncertainty
Cao et al. [68]	Tracking drift, object loss, inadequate robustness	Siamese Network (double template), MobileNet V2	Higher tracking accuracy	Consumes more training efforts
Zhao et al. [69]	Interference of cluttered background	Siamese Network, Pyramid Correlation Mapping, Channel/Spatial attention	Higher success rate of tracking	Time consuming classification
Du et al. [70]	Issues in CFT in real time	HoG, discriminative adaptive model, target re-detection	Simplified architectural implementation	Cannot be applied for objects with consist orientation changes
Chakole et al. [71]	Anomaly detection of crowd	Correlation of optical flow	Satisfactory detection	Not benchmarked
Vrskova et al. [72]	Recognition of abnormal human activity	Long Short-Term Memory	96% of classification accuracy	Induces higher computational burden
Chang et al. [73], Yahaya et al. [74], Yang et al. [75]	Anomaly detection	k-means, autoencoder, support vector machine, deep learning	Multi-target anomaly detection	Some of the outliers are also subjected to training
Rehman et al. [76]	Anomaly detection	Multi-modal (visual and acoustic source)	Improved accuracy.	Restricted to limited data, no extensive analysis over broader dataset size
Benegui and Ionescu [77]	Authentication using difference sources	Deep Neural Network, Support Vector Machine	Higher accuracy score	Not much suitable for larger dataset
Shin et al. [78]	Representation of multimodal data for anomaly detection	Analytical model using temperature, elevation changes, pattern-specific behavior of human	Ideal for surveillance map in smart city	Study not benchmarked
Dave et al. [79]	Privacy preservation during recognition of human activity	Self-supervised learning	Non-dependency of labelling	Model restricted to specific use-case only
He et al. [80]	Privacy preservation	Object swapping and blurring	Retains better accuracy and offers imperceptibility	Not proven using complex forms of object
Zhang et al. [81]	Privacy preservation from visual object detection	Blockchain and federated learning	Offers potential security from maximal threats	Higher cost of implementation
Bai et al. [82]	Privacy preservation	Multiple secret shares, CNN	Offers multi-stage detection, and optimal privacy	Higher resource dependencies

IV. Dataset and Performance Metric

After reviewing various approaches and research models, the dataset reportedly uses different forms. The frequently used dataset is *ImageNet VID* [83]. It is a benchmarked dataset with training videos of 3862 and validation videos of 555. The frame rates are maintained at 20-30 frames per second while an annotation is provided on all the Videos. Another commonly used dataset is EPIC KITCHEN, which has predefined 290 classes of an object with bounding boxes present in video samples of 32 kitchens [84]. There are also standard datasets used for specific purposes in VODT, viz. i) generalized object detection from the PASCAL dataset [85][86], ii) pedestrian data from Caltech [87], iii) indexing and retrieval of Video from TRECVID [88], iv) categorization data of human activities from HMDB-51 [89], v) annotated and segmented data of sports from Sports-1M [90], vi) tracking of an object from the MOT dataset [91] and VOT dataset [92], vii) detection of a mobile object from the CDnet2014 dataset [93], and viii) segmentation of an object from the DAVIS dataset [94]. Irrespective of availability of wider ranges of dataset, there is considerably a smaller number of research implementation towards involving multi-modal approaches of VODT.

Various performance metrics are deployed while assessing the effectiveness of VODT; however, they are all connected with accuracy-based attributes. The most common performance metric is *mean Average Precision* (mAP), used for assessing an object's traditional detection and tracking. The metric mAP is associated with accuracies of classification and regression. Before evaluating using mAP, the existing system also evaluates confidence scores and true positives. Based on the speed of the mobile object, the inference of mAP is specified in the form of fast, medium, and slow. Certain discussion states that mAP cannot be solely used for performance evaluation as it cannot capture its temporal attributes [95]. Hence, another metric termed *average delay* has surfaced, which computes the number of video frames considered for Detection and tracking from the initial frame. A dataset named ImageNet VIDT was used to verify the appropriateness of the average delay. The study outcome states that satisfactory average precision can be confirmed if the method is witnessed to maximize the average delay value. On the contrary, the maximized value of average delay will also signify maximized delay in Detection and tracking performance. Therefore, it can be stated that if any method deploys average precision as a sole performance metric, then it is challenging to evaluate the actual average delay score. It will eventually conclude that average precision

is insufficient to represent temporal attributes of the assessing framework for VODT. Hence, it is suggested to fine-tune the performance metric based on the research problem undertaken in the study model. The next section discusses the analysis of the research trend.

V. Research Trend Analysis

From the prior sections, different methodology variants are being deployed toward VODT. However, it is also noted that various common methodologies are being used to consider different use cases of VODT. It is noted that use cases play a critical role in classifying VODT methodology owing to the inclusion of discreet challenges and characteristics. Hence, to understand research trends, the primary assessment has been carried out to identify the popularity of consideration of such use cases in research publications. Fig. 1 highlights the trends of publication of such use-case adoption in VODT.

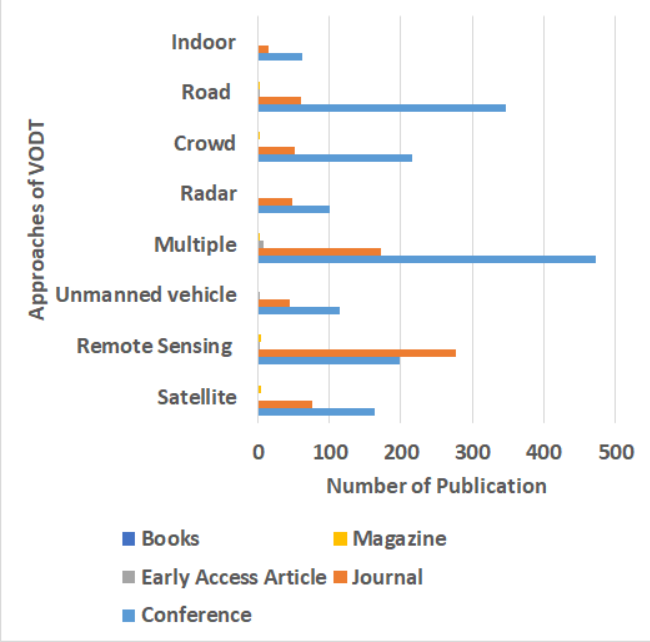

Fig. 1. Trends towards research publication.

The observation towards arriving at the graphical outcome in Fig. 1 is carried out from several different research papers published between 2017 to 2022 from IEEE Xplore, Springer, Elsevier, Wiley, ACM, MDPI, etc. The total number of count of papers for the allocated use case is more than the exhibited figure. This is because the data for Fig. 1 is obtained by filtering out only the significant experiments, excluding all the discussion or conceptual-based publications. The idea was to grab the information associated with including prominent methodologies involved in solving challenges of use cases in VODT. From this graphical outcome of the trend, it is noticed that a greater number of studies are concentrated on remote Sensing and multiple object detection method. Nearly an equivalent number of studies are on satellite-based approaches, unmanned vehicle-based VODT, radar-based VODT, and road/crowd-based VODT. Studies towards indoor-based VODT are significantly fewer. The analysis is also carried out to identify trends in methodologies by removing the constraint of year-based publication. It is noted that evolving trends for VODT are flow-based, LSTM-based, attention-network-based, tracking-based, and miscellaneous approaches, as stated in Fig. 2.

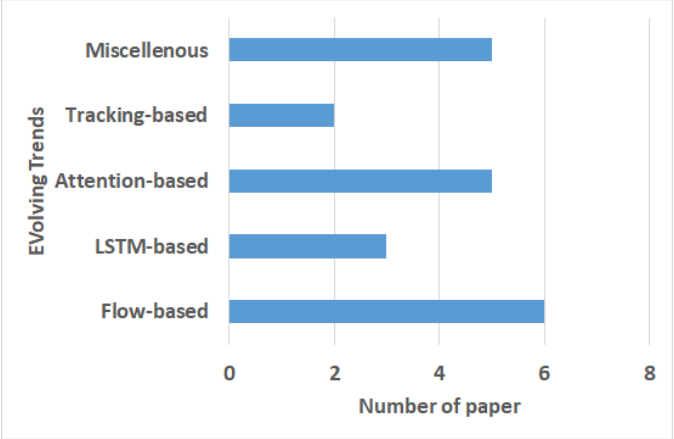

Fig. 2. Observation of evolving trend.

A closer look into multiple evolving strategies in Fig. 2 concludes that the VODT mechanism is broadly classified concerning feature aggregation and propagation. *Flow-based mechanism* adopts optical flows in dual directions. Significant contribution work was noted towards adopting this technique, viz. Deep Feature Flow [96], feature aggregation with guided flow [97], impression network [98][99][100], flow algorithm using the difference between adjacent frames and adoption of time and spatial factor of the frame, deep learning over FlowNet [101]. Further, adopting *Short-Term Long Memory* (LSTM) is another evolving trend toward maximum utilization of both time and spatial factors associated with video frames. Some of the significant contributions have been carried out by introducing unique techniques, e.g., convolution LSTM [102], online [103], and offline LSTM methods [104] [105]. *The attention-based technique* can analyze the long-duration Video for aligning the feature map with a target to minimize dependency on computational resources. The implication of such mechanism was noted in the form of various approaches viz. approaches considering only temporal factors locally [106] and global factors [107], hybrid method integrating both local and global time-based factors [108], regression and classification-based fusion of feature maps [109], managing external memory with guided objected for global aggregation [107]. The next technique is *tracking-based*, where temporal information is utilized for object detection over video frames with a fixed interval. Some of the significant approaches noted for this approach are adaptive frame-based tracking [110][111], construction of forward and backward trackers [112], adoption of refinement network for integrated Detection and tracking [113], and convolution network-based tracking [114]. Apart from the above-mentioned evolving trends of approaches, there are also *miscellaneous* methods [56]-[66]. The analysis of this approach is further carried out from the perspective of time-based evolution. Fig. 3 highlights the year-wise proposition of the above-stated approaches. The adoption of the ImageNet VID dataset [83] was the first to be evolved in 2015, followed by the evolution of tracking and Detection with cooperation TDC [112] in 2016. There has been progressively developed in the year 2017, which witnessed the Flow of Deep feature FDF

[115] for the first time along with feature aggregation with guided flow FAGF [107] and impression network [100].

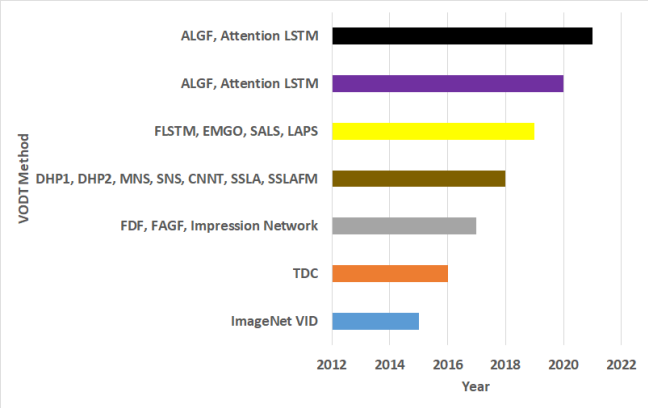

Fig. 3. Occurrence of evolving trends year wise.

The speed of new techniques further multiplied in 2018, where multiple techniques have been reported, viz. Detection with high-performance DHP1 [99] and DHPP2 [98] sampling network with spatiotemporal SNS [116], Memory network with spatiotemporal MNS [117], Convolution Neural Network based Tubelets CNNT [118], Detection of a single shot with LSTM and attention SSLA [119], Detection of a single shot with LSTM and attention with feature map LSTM-SSLAFM [120], etc. A similar pace of research trend was reported in the year 2019, where flow and LSTM have potentially improved along with new techniques of Relation Distillation Network FLSTM [106], external memory with guided object EMGO [107], semantic aggregation of level sequence SALS [109], and local attention with progressive sparsity LAPS [121]. The year 2020 and 2021 has witnessed growth in the aggregation of local and global features with improved memory ALGF [108] along with attention LSTM.

Fig. 4 showcases the effectiveness of evolving trends concerning precision scores considering post-precision (PP) and without post-precision. The outcome shows that the involvement of PP always increases the precision, which directly affects the accuracy performance. The score is found to be good for tracking-based and attention-based approaches.

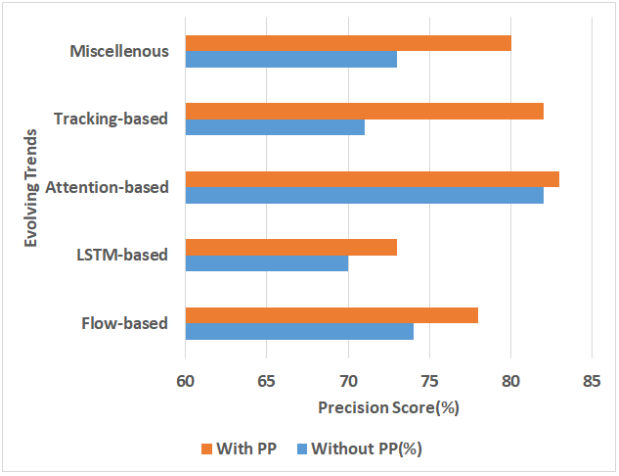

Fig. 4. Precision score for evolving trend.

VI. Result and Discussion

This section presents discussion about gist of learning inference from current review work as well as it also presents a vivid discussion associated with open-end issues of considered methodologies in VODT.

A. Discussion of Current Review

From the prior section, it is noted that there is varied implementation scheme presented towards VODT with unique agenda. One thing is quite clear that every individual research approaches deals with unique set of problems and distinct use-cases. On the other hand, proposed scheme presents review discussion about existing methodologies with an agenda to find its strength and weakness. Hence, it is infeasible to do comparison between implementation-based work and review-based work. However, the study hypothesizes that every research implementation paper must offer a clear-cut highlight of their applicability, about their strength, and about their weakness. Hence, the proposed review considers this common factor to do comparison in order to understand the degree of information associated with merits and demerits linked with suggestion towards further improving VODT. The comparison of the proposed study shown in Table II with the existing considered papers are as follows:

TABLE II. Comparing Proposed System with Considered Approaches

Approaches	Merit	Demerit
Satellite-based VODT [29]-[37]	Emphasize mechanism to increase accuracy	Doesn't present plan towards increasing adaptivity
Remote Sensing based VODT [38]-[45]	Presents technique to improve detection rate	No discussion towards extensive scope of analysis
Unmanned VODT [46]-[55]	Discusses method to increases accuracy	No solution for increased response time
Real-Time object tracking [67]-[70]	Presents higher stability modelling	No highlights towards addressing less consistent outcomes
Behavioual Analysis/Event Detection [71]-[75]	Supportive of Multi-target detection	No discussion towards increased computational complexity
Data source integration [76]-[78]	Targets higher accuracy	Lack of benchmarking
Privacy & Ethics [79]-[82]	Targets higher security	Higher cost of implementation
Proposed Study	Showcase indicators to balance accuracy and computational complexity	Overlooks security improvements

From Table II, it can be seen that proposed review manuscript offers more information associated with the applicability of various methodologies as well as highlights of various identified issues (discussed in next sub-section) which contributes towards balancing the accuracy demands with computational efficiency demands that is found majorly lacking in maximum considered methodologies in VODT. Further position of current review work with the existing review work is showcased in Table III.

TABLE III. POSITION OF CURRENT REVIEW WITH EXISTING REVIEW

Approaches	Merit	Demerit
Kaur & Singh [3]	Highlight of essential approaches	Emphasize only on deep learning approach
Salari et al. [8]	Informative contents towards scope and challenges in VODT	Lacks highlights of exhaustive research gap discussion
Tulbure et al. [21]	Comprehensive discussion about approaches	Emphasize only on deep CNN approaches
Zhang et al. [29]	Presents methods to deal with complex image-based approaches	Discussion restricted to satellite videos
Proposed Study	Captures all taxonomies of VODT, presents clear highlights of trends, clear cut representation of current research gap	Inference carried out for recent implementation study only

B. Research Gap Analysis

After reviewing the existing schemes of VODT, it is noted that many open-end challenges are still associated with the context of methodologies and use cases considered A closer look into the efficiencies of all the model exhibit that there is still an open-end issue mainly associated with the speed and accuracy factors balancing with the computational complexities and resource demands. These are some of the areas that have been less emphasized upon. Apart from this, there are few benchmarked datasets which consists of each framed to be labelled. Further, it is noted that ImageNet VID, which is one of the frequently used datasets in VODT, doesn't possess the practical complexities of real scene. Further, the dataset consists of only few objects which are not recommended to be investigated for complex VODT application design. Further, existing schemes either considers global or local information associated with temporal attributes separately. This section outlines the prominent issues being identified in the form of a research gap:

- Issues with the VODT dataset: There are many available datasets for performing VODT, but they lack major benchmarking. The available video dataset doesn't have challenges like a practical environment. Apart from this, very limited objects are present in each frame. Hence, accuracy will always be anticipated to be better while using such video datasets; however, they are less likely to apply to real-world applications.
- Issues with performance metrics: There is no specific standard performance metric for VODT analysis. It should be noted that the performance metric of mAP is extracted from the experiments from object detection of an image. However, mAP cannot process temporal attributes, leading to the evolution of average delay with this evaluation capability. However, this metric cannot evaluate the stability factor associated with the video dataset. Hence, varied ranges of accuracy-based metrics are in practice for assessing the effectiveness of VODT.
- Narrowed Scope of Optimization: An algorithm must encapsulate the maximum constraint condition to optimize detection and track performance accuracy. A better form of constraint modeling can harness maximum information from the context of Video in the dataset and its associated dynamicity. It demands the utilization of local and global information; however, very few reported studies have such consideration. This imposes an extensive challenge to implement the optimization of VODT.
- Adoption of Complex Method: It is noted that learning-based algorithms are extensively used to improve the overall performance of VODT in many cases. However, such learning algorithms are associated with higher ranges of complexities that are not found to be addressed in existing studies. Adopting such a method (e.g., Convolution Neural Network, Siamese Network) has higher dependencies on training data, indirectly increasing the dependencies of resources required to process it. As a result, it also results in slower computation speed.
- The trade-off between research and practical demand: Although various studies use real-time datasets or developed hardware for VODT, none of them have been analyzed in a real-time environment in the practical world. It is also noted that attention-based approaches offer higher accuracy; however, they have higher processing demand, which impedes catering to real-time demands.
- Other critical challenges: The existing studies carried out towards event detection don't deal with data variability as well as scalability. Further, studies considering recognition of human activity and detection of anomalies are highly dependent on orientation and viewpoints which always vary and such challenges are not addressed in object detection. Identification of real-time object detection and tracking are also quite challenging to be accomplished because of same reason.

From the outcome of the above-mentioned research gap associated with existing video object detection methodologies, it is noted that existing schemes do have their own advantages as well as significant limiting attributes. However, from a global context viewpoint, there are certain contributory suggestions which the proposed study has arrived towards improving or addressing the above-mentioned research gaps. Following are some contributory suggestions towards leveraging the performance of existing video object detection methodologies:

- Encouraging consideration of diverse and large-scale training data: Different challenges associated with backgrounds and wide ranges of classes of an object can be effectively addressed considering diverse dataset with better representation. The idea is to achieve better performance on real-time events and effective generalization.
- Need focus on architecture design: Majority of existing studies are use-case specific with not much emphasizing over architecture-based deployment and extending its applicability. The issues of computational

burden and accuracy can be well balanced if architectures are subjected to optimization.

- Needs practical predictive approach: There are various existing predictive models using branches of artificial intelligence (machine learning and deep learning). However, sustainability of such model over extensive test-cases is not verified. Hence, there is a need of a learning model that can adapt to dynamically changing condition is highly demanded. More research towards ensemble approaches can offer extended robustness and minimize system biases.
- Inclusion of contextual information: The problems associated with accuracy of localization of an object as well as uncertainty connected with instances of an object can be well addressed by including contextual information. Extraction of contextual information can be carried out from semantic segmentation, understanding the scene, and also from surrounding objects.

Hence, all the above-mentioned points are contributory suggestion towards improving video object detection methodology. The next section summarizes the essential highlights of this review work.

VII. Conclusion

This paper offers an overview of some notable techniques and approaches towards the VODT, which plays one of the crucial intrinsic operations within any surveillance system. Prior to draw conclusive remarks, it is to be noted that proposed review work formulates certain research questions prior to undertake the review work as follows:

- Ro1: Is the existing review paper towards VODT informative enough to draw conclusive remarks about strength and weakness?
- Ro2: What are the frequently used approaches towards VODT?
- Ro3: What are the issues in multiple approaches towards VODT?
- Ro4: How to improve the performance of VODT differently from existing studies.

After reviewing the existing studies both on perspective of implementation and review work, it is noted that existing papers cannot be used for withdrawing conclusive remarks as they are highly symptomatic in nature. It will mean that they deal with narrowed set of issues while leaving other associated issues unaddressed. This is the response towards Ro_1. The response to Ro_2 is that attention-based and tracking-based approaches are most effectively proven and hence frequently adopted. Towards the response for Ro_3, the multiples issues have been presented in the form of research gap discussed in prior section. The similar sub-section of research gap are also essential points that are required to be considered to improve the performance of VODT that are not presented in existing studies. This acts as response for Ro_4. After reviewing various approaches, it is noted that approaches toward VODT are highly specific to the use cases. The studies are mainly not emphasized in the methodologies. At the same time, more inclination of existing approaches is found towards solving the problems associated with unique use cases, e.g., satellite-based VODT, unmanned aerial vehicle-based VODT, remote sensing-based studies, multiple tracking-based studies, etc. A smaller number of generalized frameworks can address all the issues about high performance and robust tracking and Detection. The analysis also found that the implication of deep learning, Siamese network, and convolution neural networks is constantly rising to improve Detection and tracking performance. Looking into the progress in the study, it can be concluded that there is a long way to go to witness a more high-performance VODT approach. There is a need for more extensive analysis to expose the VODT approach to challenging scenarios of the practical world. In contrast, the existing approaches are too narrow and confined to their adopted research environment. The contribution and novelty of this paper is

1) Unlike existing review work, the current paper captures maximum deployment area, use-cases, and wider variants methodologies adopted towards improving performance of VODT.

2) This paper discusses the new classification of approaches of VODT concerning discreet use cases not reported in existing review work,

3) the paper also introduces a compact discussion associated with the commercial application of VODT, where it can be seen that progress made by commercial application and research work has a wide gap,

4) the paper discusses all the notable research contributions concerning the problems being addressed, the methodology being adopted, their respective strength and weakness,

5) the paper presents a discussion of research trends of VODT to find that evolving approaches and their frequencies of usage,

6) the paper outlines the research gap to signify the importance of enhancing the existing studies for developing high-performance VODT.

The future work will be in the direction of addressing the identified research gap from this review work. For this purpose, the initial work direction will be developing a robust detection module considering the challenging scene context of the video feed. Upon accomplishing a satisfactory detection module assessed over different extensive test environments, the next work will be carried out towards tracking operation. As tracking operation is seamless, both spatial and temporal attributes will be considered for mathematical modeling of the VODT approach that can lead to better optimization of its performance. The sole motive of future work will be to accomplish high-performance VODT in a cost-effective computational approach.

References

[1] M. H Kolekar, Intelligent Video Surveillance Systems-An Algorithmic Approach, CRC Press, ISBN: 9781351649902, 1351649906, 2018

[2] E. Maiettini, G. Pasquale, L. Rosasco, L. Natale, "Online object detection: a robotics challenge," ACM-Autonomous Robots, Vol.44,

No.5, pp.739–757, 2020. DOI: https://doi.org/10.1007/s10514-019-09894-9

[3] B. Kaur, S. Singh, "Object Detection using Deep Learning: A Review," ACM- Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence, pp.328–334, 2021. DOI:https://doi.org/10.1145/3484824.3484889

[4] A. Boukerche, Z. Hou, "Object Detection Using Deep Learning Methods in Traffic Scenarios," ACM Computing Surveys, Vol.54, Issue 2, Article No.:30, pp.1–35, 2022. DOI: https://doi.org/10.1145/3434398

[5] Y. Ji, P. Yin, X. Sun, K. H. H. B. Ghazali, N. Guo, "A Comparative Study and Simulation of Object Tracking Algorithms," ACM-The 4th International Conference on Video and Image Processing, pp.161–167, December 2020. DOI: https://doi.org/10.1145/3447450.3447476

[6] Y. Wang, J-N Hwang, G. Wang, H. Liu, K-J Kim, H-M Hsu, J. Cai, H. Zhang, Z. Jiang, R. Gu, "ROD2021 Challenge: A Summary for Radar Object Detection Challenge for Autonomous Driving Applications", ACM-Proceedings of the 2021 International Conference on Multimedia Retrieval, pp.553–559, August 2021. DOI: https://doi.org/10.1145/3460426.3463658

[7] E. Arulprakash, M. Aruldoss, "A study on generic object detection with emphasis on future research directions," ACM-Journal of King Saud University - Computer and Information Sciences, Vol.34, Issue 9, pp 7347–7365, Oct 2022. DOI: https://doi.org/10.1016/j.jksuci.2021.08.001

[8] A. Salari, A. Djavadifar, X. Liu, H. Najjaran, "Object recognition datasets and challenges: A review," ACM-Neurocomputing, Vol.495, Issue C, pp.129–152, Jul 2022. DOI: https://doi.org/10.1016/j.neucom.2022.01.022

[9] İ. Delibaşoğlu, "Surveillance with UAV Videos", Intechopen-Intelligent Video Surveillance - New Perspectives, 2022. DOI: 10.5772/intechopen.105959

[10] H. Li, Y. Dong, L. Xu, S. Zhang, & J. Wang, "Object detection method based on global feature augmentation and adaptive regression in IoT," SpringerOpen-Neural Computing and Applications, vol.33, pp.4119–4131, 2021

[11] J-P Mercier, M. Garon, P. Giguère, J-F Lalonde, "Deep Template-based Object Instance Detection", arXiv- Computer Vision and Pattern Recognition, 2019. DOI: https://doi.org/10.48550/arXiv.1911.11822

[12] Zhu, J.; Wang, Z.; Wang, S.; Chen, S. Moving Object Detection Based on Background Compensation and Deep Learning. *Symmetry* 2020, *12*, 1965. https://doi.org/10.3390/sym12121965

[13] A. S. Patel, R. Vyas, O. P. Vyas, M. Ojha, V. Tiwari, "Motion-compensated online object tracking for activity detection and crowd behavior analysis," SpringerOpen-The Visual Computer, 2022

[14] K-P Kortmann, J. Zumsande, M. Wielitzka, T. Ortmaier, "Temporal Object Tracking in Large-Scale Production Facilities using Bayesian Estimation," Elsevier-IFAC-PapersOnLine, Vol.53, No,2, pp.11125-11131, 2022. DOI: https://doi.org/10.1016/j.ifacol.2020.12.271

[15] A. Mishra, S. Lee, D. Kim, S. Kim, "In-Cabin Monitoring System for Autonomous Vehicles", *Sensors*, vol.*22*, No.4360, 2022. DOI: https://doi.org/10.3390/s22124360

[16] V. Paidi, H. Fleyeh, J. Håkansson, and R. G. Nyberg, "Tracking Vehicle Cruising in an Open Parking Lot Using Deep Learning and Kalman Filter," Hindawi-Journal of Advanced Transportation, Article ID 1812647, DOI:https://doi.org/10.1155/2021/1812647

[17] T. Nguyen, C. Pham, K. Nguyen, M. Hoai, "Few-shot Object Counting and Detection," arXiv, Computer Vision and Pattern Recognition, 2022. DOI: https://doi.org/10.48550/arXiv.2207.10988

[18] J.S. Murthy, G. M. Siddesh, W-C Lai, B. D. Parameshachari, S.N. Patil, and K. L. Hemalatha, "ObjectDetect: A Real-Time Object Detection Framework for Advanced Driver Assistant Systems Using YOLOv5", Hindawi-Wireless Communication and Mobile Computing, Article ID 9444360, 2022, DOI:https://doi.org/10.1155/2022/9444360

[19] S. Dokania, A. H. A. Hafez, A. Subramanian, M. Chandraker, C.V. Jawahar, "IDD-3D: Indian Driving Dataset for 3D Unstructured Road Scene", arXiv:2210.12878v1 [cs.CV] 23 October 2022

[20] L. Malburg, M-P Rieder, R. Seiger, P. Klein, R. Bergmann, "Object Detection for Smart Factory Processes by Machine Learning", Elsevier-Procedia Computer Science, Vol.184, pp.581-588, 2021. DOI: https://doi.org/10.1016/j.procs.2021.04.009

[21] A-A Tulbure, A-A Tulbure, E-H Dulf, "A review on modern defect detection models using DCNNs – Deep convolutional neural networks," ScienceDirect-Journal of Advanced Research Vol.35, pp.33-48, 2022. DOI: https://doi.org/10.1016/j.jare.2021.03.015

[22] L. Zhou, L. Zhang, N. Konz, "Computer Vision Techniques in Manufacturing", TechRxiv Preprint, 2021. DOI: https://doi.org/10.36227/techrxiv.17125652.v2

[23] V. Isailovic, A. Peulic, M. Djapan, M. Savkovic, A. M. Vukicevic, "The compliance of head-mounted industrial PPE by using deep learning object detectors," Scientific Reports, vol.12, Article number: 16347, 2022.

[24] J. Wen, T. Abe, and T. Suganuma, "A Customer Behavior Recognition Method for Flexibly Adapting to Target Changes in Retail Stores," Sensors, vol. 22, no. 18, p. 6740, Sep. 2022, doi: 10.3390/s22186740.

[25] Y.-S. Yoo, S.-H. Lee, and S.-H. Bae, "Effective Multi-Object Tracking via Global Object Models and Object Constraint Learning," Sensors, vol. 22, no. 20, p. 7943, Oct. 2022, doi: 10.3390/s22207943

[26] J. M. R. Andaur, G. A. Ruz, and M. Goycoolea, "Predicting Out-of-Stock Using Machine Learning: An Application in a Retail Packaged Foods Manufacturing Company," Electronics, vol. 10, no. 22, p. 2787, Nov. 2021, doi: 10.3390/electronics10222787.

[27] K. Xia et al., "An Intelligent Self-Service Vending System for Smart Retail," Sensors, vol. 21, no. 10, p. 3560, May 2021, doi: 10.3390/s21103560.

[28] E. Maltezos et al., "A Video Analytics System for Person Detection Combined with Edge Computing," computation, vol. 10, no. 3, p. 35, Feb. 2022, doi: 10.3390/computation10030035

[29] Z. Zhang, C. Wang, J. Song, and Y. Xu, "Object Tracking Based on Satellite Videos: A Literature Review," Remote Sensing, vol. 14, no. 15, p. 3674, Jul. 2022, doi: 10.3390/rs14153674.

[30] S. Chen et al., "Vehicle Tracking on Satellite Video Based on Historical Model," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 7784-7796, 2022, doi: 10.1109/JSTARS.2022.3195522.

[31] Z. Hu, D. Yang, K. Zhang, and Z. Chen, "Object Tracking in Satellite Videos Based on Convolutional Regression Network With Appearance and Motion Features," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 783-793, 2020, doi: 10.1109/JSTARS.2020.2971657.

[32] F. Shi, F. Qiu, X. Li, Y. Tang, R. Zhong, and C. Yang, "A Method to Detect and Track Moving Airplanes from a Satellite Video," Remote Sensing, vol. 12, no. 15, p. 2390, Jul. 2020, doi: 10.3390/rs12152390.

[33] D. Wu, H. Song, and C. Fan, "Object Tracking in Satellite Videos Based on Improved Kernel Correlation Filter Assisted by Road Information," Remote Sensing, vol. 14, no. 17, p. 4215, Aug. 2022, doi: 10.3390/rs14174215.

[34] S. Xuan, S. Li, M. Han, X. Wan and G. -S. Xia, "Object Tracking in Satellite Videos by Improved Correlation Filters With Motion Estimations," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 2, pp. 1074-1086, Feb. 2020, doi: 10.1109/TGRS.2019.2943366.

[35] Y. Zhang, D. Chen, and Y. Zheng, "Satellite Video Tracking by Multi-Feature Correlation Filters with Motion Estimation," Remote Sensing, vol. 14, no. 11, p. 2691, Jun. 2022, doi: 10.3390/rs14112691.

[36] Z. Zhou, S. Li, W. Guo, and Y. Gu, "Few-Shot Aircraft Detection in Satellite Videos Based on Feature Scale Selection Pyramid and Proposal Contrastive Learning," Remote Sensing, vol. 14, no. 18, p. 4581, Sep. 2022, doi: 10.3390/rs14184581

[37] K. Zhu et al., "Single Object Tracking in Satellite Videos: Deep Siamese Network Incorporating an Interframe Difference Centroid Inertia Motion Model," Remote Sensing, vol. 13, no. 7, p. 1298, Mar. 2021, doi: 10.3390/rs13071298.

[38] Y. You, J. Cao, and W. Zhou, "A Survey of Change Detection Methods Based on Remote Sensing Images for Multi-Source and Multi-Objective Scenarios," Remote Sensing, vol. 12, no. 15, p. 2460, Jul. 2020, doi: 10.3390/rs12152460.

[39] L. Lei and D. Guo, "Multitarget Detection and Tracking Method in Remote Sensing Satellite Video," Hindawi-Computational Intelligence

and Neuroscience, Article ID 7381909, 2021. DOI:https://doi.org/10.1155/2021/7381909

[40] Q. Lin, "Real-Time Multitarget Tracking for Panoramic Video Based on Dual Neural Networks for Multisensor Information Fusion," Hindawi-Mathematical Problems in Engineering, Article ID 8313471, 2022. DOI:https://doi.org/10.1155/2022/8313471

[41] T.J. Ma, "Remote sensing detection enhancement," Springer-Journal of Big Data, vol.8, No.127, 2021.DOI: https://doi.org/10.1186/s40537-021-00517-8

[42] G. Tochon, J. Chanussot, M. Dalla Mura and A. L. Bertozzi, "Object Tracking by Hierarchical Decomposition of Hyperspectral Video Sequences: Application to Chemical Gas Plume Tracking," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 8, pp. 4567-4585, Aug. 2017, doi: 10.1109/TGRS.2017.2694159.

[43] B. Uzkent, M. J. Hoffman and A. Vodacek, "Integrating Hyperspectral Likelihoods in a Multidimensional Assignment Algorithm for Aerial Vehicle Tracking," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 9, pp. 4325-4333, Sept. 2016, doi: 10.1109/JSTARS.2016.2560220.

[44] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 1, pp. 449-461, Jan. 2019, doi: 10.1109/TGRS.2018.2856370.

[45] J. Wei, J. Sun, Z. Wu, J. Yang, and Z. Wei, "Moving Object Tracking via 3-D Total Variation in Remote-Sensing Videos," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 3506405, doi: 10.1109/LGRS.2021.3077257.

[46] E. Çintaş, B. Özyer, and E. Şimşek, "Vision-Based Moving UAV Tracking by Another UAV on Low-Cost Hardware and a New Ground Control Station," in IEEE Access, vol. 8, pp. 194601-194611, 2020, doi: 10.1109/ACCESS.2020.3033481.

[47] C. Deng, S. He, Y. Han, and B. Zhao, "Learning Dynamic Spatial-Temporal Regularization for UAV Object Tracking," in IEEE Signal Processing Letters, vol. 28, pp. 1230-1234, 2021, doi: 10.1109/LSP.2021.3086675.

[48] Z. Ding, S. Liu, M. Li, Z. Lian, and H. Xu, "A Blockchain-Enabled Multiple Object Tracking for Unmanned System With Deep Hash Appearance Feature," in IEEE Access, vol. 9, pp. 1116-1123, 2021, doi: 10.1109/ACCESS.2020.3046243.

[49] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1758-1770, June 2020, doi: 10.1109/TCSVT.2019.2905881.

[50] F. Lin, C. Fu, Y. He, F. Guo, and Q. Tang, "Learning Temporary Block-Based Bidirectional Incongruity-Aware Correlation Filters for Efficient UAV Object Tracking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 6, pp. 2160-2174, June 2021, doi: 10.1109/TCSVT.2020.3023440.

[51] Y. Bühler, L. Meier and C. Ginzler, "Potential of Operational High Spatial Resolution Near-Infrared Remote Sensing Instruments for Snow Surface Type Mapping," in IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 4, pp. 821-825, April 2015, doi: 10.1109/LGRS.2014.2363237.

[52] Y. Shan, S. Liu, Y. Zhang, M. Jing, and H. Xu, "LMD-TShip⋆: Vision Based Large-Scale Maritime Ship Tracking Benchmark for Autonomous Navigation Applications," in IEEE Access, vol. 9, pp. 74370-74384, 2021, doi: 10.1109/ACCESS.2021.3079132.

[53] X. Xue, Y. Li, X. Yin, C. Shang, T. Peng, and Q. Shen, "Semantic-Aware Real-Time Correlation Tracking Framework for UAV Videos," in IEEE Transactions on Cybernetics, vol. 52, no. 4, pp. 2418-2429, April 2022, doi: 10.1109/TCYB.2020.3005453.

[54] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization," in IEEE Transactions on Industrial Electronics, vol. 69, no. 6, pp. 6004-6014, June 2022, doi: 10.1109/TIE.2021.3088366.

[55] Q. Yu, B. Wang, and Y. Su, "Object Detection-Tracking Algorithm for Unmanned Surface Vehicles Based on a Radar-Photoelectric System," in IEEE Access, vol. 9, pp. 57529-57541, 2021, doi: 10.1109/ACCESS.2021.3072897.

[56] S. Banerjee, H. H. Chopp, J. G. Serra, H. T. Yang, O. Cossairt, and A. K. Katsaggelos, "An Adaptive Video Acquisition Scheme for Object Tracking and Its Performance Optimization," in IEEE Sensors Journal, vol. 21, no. 15, pp. 17227-17243, 1 August 1, 2021, doi: 10.1109/JSEN.2021.3081351.

[57] J. Chen, H. -W. Huang, P. Rupp, A. Sinha, C. Ehmke, and G. Traverso, "Closed-Loop Region of Interest Enabling High Spatial and Temporal Resolutions in Object Detection and Tracking via Wireless Camera," in IEEE Access, vol. 9, pp. 87340-87350, 2021, doi: 10.1109/ACCESS.2021.3086499.

[58] L. Cheng, J. Wang, and Y. Li, "ViTrack: Efficient Tracking on the Edge for Commodity Video Surveillance Systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 3, pp. 723-735, 1 March 2022, doi: 10.1109/TPDS.2021.3081254.

[59] H. Gu et al., "A Collaborative and Sustainable Edge-Cloud Architecture for Object Tracking with Convolutional Siamese Networks," in IEEE Transactions on Sustainable Computing, vol. 6, no. 1, pp. 144-154, 1 Jan.-March 2021, doi: 10.1109/TSUSC.2019.2955317.

[60] S. J. Lee, S. Lee, S. I. Cho, and S. -J. Kang, "Object Detection-Based Video Retargeting With Spatial–Temporal Consistency," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4434-4439, Dec. 2020, doi: 10.1109/TCSVT.2020.2981652.

[61] C. -J. Liu and T. -N. Lin, "DET: Depth-Enhanced Tracker to Mitigate Severe Occlusion and Homogeneous Appearance Problems for Indoor Multiple-Object Tracking," in IEEE Access, vol. 10, pp. 8287-8304, 2022, doi: 10.1109/ACCESS.2022.3144153.

[62] M. R. Marshall et al., "3-D Object Tracking in Panoramic Video and LiDAR for Radiological Source–Object Attribution and Improved Source Detection," in IEEE Transactions on Nuclear Science, vol. 68, no. 2, pp. 189-202, Feb. 2021, doi: 10.1109/TNS.2020.3047646.

[63] R. Mostafa, H. Baraka, and A. Bayoumi, "LMOT: Efficient Light-Weight Detection and Tracking in Crowds," in IEEE Access, vol. 10, pp. 83085-83095, 2022, doi: 10.1109/ACCESS.2022.3197157.

[64] B. Ramesh et al., "e-TLD: Event-Based Framework for Dynamic Object Tracking," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3996-4006, Oct. 2021, doi: 10.1109/TCSVT.2020.3044287.

[65] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-Counting: Using Network Flows on Crowd Density Maps for Tracking Multiple Targets," in IEEE Transactions on Image Processing, vol. 30, pp. 1439-1452, 2021, doi: 10.1109/TIP.2020.3044219.

[66] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep Affinity Network for Multiple Object Tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 104-119, 1 January 2021, doi: 10.1109/TPAMI.2019.2929520.

[67] H. Zhang and X. Ren, "Robust real-time object tracking system for human-following quadruped robot," in *Lecture Notes in Electrical Engineering*, Singapore: Springer Nature Singapore, 2022, pp. 388–397

[68] J. Cao *et al.*, "Robust object tracking algorithm for autonomous vehicles in complex scenes," *Remote Sens. (Basel)*, vol. 13, no. 16, p. 3234, 2021, doi: 10.3390/rs13163234

[69] W. Zhao, M. Deng, C. Cheng, and D. Zhang, "Real-time object tracking algorithm based on Siamese network," *Appl. Sci. (Basel)*, vol. 12, no. 14, p. 7338, 2022, doi: 10.3390/app12147338

[70] C. Du, M. Lan, M. Gao, Z. Dong, H. Yu, and Z. He, "Real-time object tracking via adaptive correlation filters," *Sensors (Basel)*, vol. 20, no. 15, p. 4124, 2020, doi: 10.3390/s20154124.

[71] P. D. Chakole, V. R. Satpute, and N. Cheggoju, "Crowd behavior anomaly detection using correlation of optical flow magnitude," *J. Phys. Conf. Ser.*, vol. 2273, no. 1, p. 012023, 2022, doi: 10.1088/1742-6596/2273/1/012023.

[72] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "A new approach for abnormal human activities recognition based on ConvLSTM architecture," *Sensors (Basel)*, vol. 22, no. 8, 2022, doi: 10.3390/s22082946.

[73] Y. Chang *et al.*, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognit.*, vol. 122, no. 108213, p. 108213, 2022, doi: 10.1016/j.patcog.2021.108213

[74] S. W. Yahaya, A. Lotfi, and M. Mahmud, "Detecting anomaly and its sources in activities of daily living," *SN Comput. Sci.*, vol. 2, no. 1, 2021, doi: 10.1007/s42979-020-00418-2.

[75] Y. Yang, F. Angelini, and S. M. Naqvi, "Pose-driven human activity anomaly detection in a CCTV-like environment," *IET Image Process.*, vol. 17, no. 3, pp. 674–686, 2023, doi: 10.1049/ipr2.12664.

[76] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, "Multi-modal anomaly detection by using audio and visual cues," *IEEE Access*, vol. 9, pp. 30587–30603, 2021, doi: 10.1109/access.2021.3059519

[77] C. Benegui and R. T. Ionescu, "Improving the authentication with built-in camera protocol using built-in motion sensors: A deep learning solution," *arXiv [cs.CR]*, 2021. doi: 10.3390/1010000.

[78] H. Shin, K.-I. Na, J. Chang, and T. Uhm, "Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security," *ETRI J.*, vol. 44, no. 2, pp. 183–193, 2022, doi: 10.4218/etrij.2021-0395.

[79] I. R. Dave, C. Chen, and M. Shah, "SPAct: Self-supervised privacy preservation for action recognition," *arXiv [cs.CV]*, 2022. Accessed: Jun. 24, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Dave_SPAct_Self-Supervised_Privacy_Preservation_for_Action_Recognition_CVPR_2022_paper.pdf

[80] P. He *et al.*, "Privacy-preserving object detection," *arXiv [cs.CV]*, 2021. Accessed: Jun. 24, 2023. [Online]. Available: http://arxiv.org/abs/2103.06587

[81] J. Zhang, J. Zhou, J. Guo, and X. Sun, "Visual object detection for privacy-preserving federated learning," *IEEE Access*, vol. 11, pp. 33324–33335, 2023, doi: 10.1109/access.2023.3263533.

[82] T. Bai, S. Fu, and Q. Yang, "Privacy-preserving object detection with secure convolutional neural networks for vehicular edge computing," *Future Internet*, vol. 14, no. 11, p. 316, 2022, doi: 10.3390/fi14110316.

[83] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, "ImageNet Large Scale Visual Recognition Challenge", Int. J. Comput. Vis., vol.115, pp.211–252, 2015

[84] D. Damen *et al.*, "The EPIC-KITCHENS dataset: Collection, challenges and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4125–4141, 2021

[85] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[86] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[87] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: A benchmark," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 304-311, doi: 10.1109/CVPR.2009.5206631.

[88] G. Awad *et al.*, "TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking," *Nist.gov*. [Online]. Available: https://www-nlpir.nist.gov/projects/tvpubs/tv17.papers/tv17overview.pdf. [Accessed: 28-Apr-2023].

[89] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," *2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2556-2563, doi: 10.1109/ICCV.2011.6126543.

[90] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223

[91] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv [cs.CV]*, 2015

[92] M. Kristan *et al.*, "The Visual Object Tracking VOT2015 Challenge Results," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, 2015, pp. 564-586, doi: 10.1109/ICCVW.2015.79.

[93] Y. Wang, P. -M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, 2014, pp. 393-400, doi: 10.1109/CVPRW.2014.126.

[94] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[95] H. Mao, X. Yang, and W. J. Dally, "A delay metric for video object detection: What average precision fails to tell," *arXiv [cs.CV]*, 2019.

[96] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," *arXiv [cs.CV]*, 2016.

[97] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," *arXiv [cs.CV]*, 2017.

[98] X. Zhu, J. Dai, X. Zhu, Y. Wei, and L. Yuan, "Towards high performance video object detection for mobiles," *arXiv [cs.CV], 2018*

[99] F. He, N. Gao, J. Jia, X. Zhao, and K. Huang, "QueryProp: Object query propagation for high-performance video object detection," *Proc. Conf. AAAI Artif. Intell.*, vol. 36, no. 1, pp. 834–842, 2022.

[100] C. Hetang, H. Qin, S. Liu, and J. Yan, "Impression Network for video object detection," *arXiv [cs.CV]*, 2017.

[101] A. Dosovitskiy *et al.*, "FlowNet: Learning Optical Flow with Convolutional Networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 2758-2766, doi: 10.1109/ICCV.2015.316.

[102] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv [cs.LG]*, 2015.

[103] C. Zhang and J. Kim, "Modeling Long- and Short-Term Temporal Context for Video Object Detection," *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 71-75, doi: 10.1109/ICIP.2019.8802920.

[104] M. Liu and M. Zhu, "Mobile video object detection with temporally-aware feature maps," *arXiv [cs.CV]*, 2017.

[105] M. Liu, M. Zhu, M. White, Y. Li, and D. Kalenichenko, "Looking fast and slow: Memory-guided mobile video object detection," *arXiv [cs.CV]*, 2019.

[106] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Relation Distillation Networks for video object detection," *arXiv [cs.CV]*, 2019.

[107] H. Deng *et al.*, "Object Guided External Memory Network for Video Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 6677-6686, doi: 10.1109/ICCV.2019.00678.

[108] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," *arXiv [cs.CV]*, 2020.

[109] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence Level Semantics Aggregation for video object detection," *arXiv [cs.CV]*, 2019.

[110] W. Yang, B. Liu, W. Li and N. Yu, "Tracking Assisted Faster Video Object Detection," *2019 IEEE International Conference on Multimedia and Expo (ICME)*, Shanghai, China, 2019, pp. 1750-1755, doi: 10.1109/ICME.2019.00301.

[111] H. Luo, W. Xie, X. Wang, and W. Zeng, "Detect or track: Towards cost-effective video object detection/tracking," *arXiv [cs.CV]*, 2018.

[112] H.-U. Kim and C.-S. Kim, "CDT: Cooperative detection and tracking for tracing multiple objects in video sequences," in *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 2016, pp. 851–86

[113] H. Mao, T. Kong, and W. J. Dally, "CaTDet: Cascaded tracked detector for efficient object detection from video," *arXiv [cs.CV]*,

[114] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," *arXiv [cs.CV]*, 2017

[115]G. Han, X. Zhang, and C. Li, "Semi-supervised DFF: Decoupling detection and feature flow for video object detectors," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[116]G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," *arXiv [cs.CV]*, 2018.

[117]F. Xiao and Y. J. Lee, "Video object detection with an aligned spatial-Temporal Memory," *arXiv [cs.CV]*,

[118]K. Kang *et al.*, "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *arXiv [cs.CV]*, 2016.

[119]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv [cs.CV]*, 2015.

[120]X. Chen, J. Yu, and Z. Wu, "Temporally Identity-Aware SSD with Attentional LSTM," *arXiv [cs.CV]*, 2018

[121]C. Guo *et al.*, "Progressive Sparse Local Attention for Video object detection," *arXiv [cs.CV]*, 2019.