VmAP: A Fair Metric for Video Object Detection

Anupam Sobti Vaibhav Mavi M Balakrishnan Chetan Arora
Indian Institute of Technology Delhi
India
anupamsobti@cse.iitd.ac.in

ABSTRACT

Video object detection is the task of detecting objects in a sequence of frames, typically, with a significant overlap in content among consecutive frames. Mean Average Precision (mAP) was originally proposed for evaluating object detection techniques in independent frames, but has been used for evaluating video based object detectors as well. This is undesirable since the average precision over all frames masks the biases that a certain object detector might have against certain types of objects depending on the number of frames for which the object is present in a video sequence. In this paper we show several disadvantages of mAP as a metric for evaluating video based object detection. Specifically, we show that: (1) some object detectors could be severely biased against some specific kind of objects, such as small, blurred, or low contrast objects, and such differences may not reflect in mAP based evaluation, (2) operating a video based object detector at the best frame based precision/recall value (high F1 score) may lead to many false positives without a significant increase in the number of objects detected. (3) mAP does not take into account that tracking can be potentially used to recover missed detections in the temporal neighborhood while this can be account for while evaluating detectors. As an alternate, we suggest a novel evaluation metric (VmAP) which takes the focus away from evaluating detections on every frame. Unlike mAP, VmAP rewards a high recall of different object views throughout the video. We form sets of bounding boxes having similar views of an object in a temporal neighborhood and use a set-level recall for evaluation. We show that VmAP is able to address all the challenges with the mAP listed above. Our experiments demonstrate hidden biases in object detectors, shows upto 99% reduction in false positives while maintaining similar object recall and shows a 9% improvement in correlation with post-tracking performance.

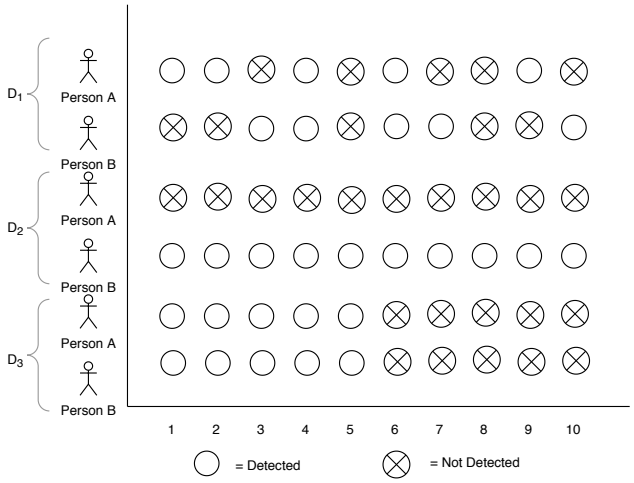


Figure 1: Consider a scenario, when two pedestrians A, and B, are walking in front of a mobile robot. Each pedestrian is visible in the camera for 10 frames. Suppose, a detector $\mathcal{D}_1$ detects each pedestrian A, and B for 5 random frames, $\mathcal{D}_2$ detects B in all frames and misses A completely, and $\mathcal{D}_3$ detects A, and B in 5 frames but only in the frames when the objects are nearer (and thus visually larger). It is easy to see that the precision of all 3 detectors is identical. However, $\mathcal{D}_1$ is most likely to detect all view variations of objects A and B in a video, $\mathcal{D}_3$ would only work when object size is large, whereas $\mathcal{D}_2$ is unfair and may even cause the robot to collide with pedestrian A. The example highlights the case where three detectors with very different characteristics (or even having a bias) may have a similar performance in terms of mAP.

CCS CONCEPTS

• **Computing methodologies → Object detection**.

KEYWORDS

video object detection, metric, evaluation, fairness, bias

ACM Reference Format:
Anupam Sobti Vaibhav Mavi M Balakrishnan Chetan Arora. 2021. VmAP: A Fair Metric for Video Object Detection. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3474085.3475383

1 INTRODUCTION

It is widely accepted that large benchmark datasets have been one of the primary reasons for rapid progress achieved in many computer vision problems [10, 15, 26]. In this progress, metrics for comparing the performance of various algorithms on these benchmark datasets have also played a critical role. These metrics enable comparison of widely different algorithmic techniques using various kinds of hyper-parameters, and bring out the strengths and weaknesses of each detection method. For example, accuracy is a good indicator of an algorithm's performance for object detection, but is often tied to a particular value of precision and recall. To overcome this dependence, researchers have been using mAP, which takes an

MM '21, October 20–24, 2021, Virtual Event, China
2021. ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475383

average precision (AP) over all recall values and further calculates the mean of AP over all classes. This has served the community well for evaluating object detectors on independent frames.

Many computer-vision based applications of interest often work with video/streaming image inputs. For example, pedestrian and vehicle detection systems in autonomous vehicles, surveillance applications, medical diagnostics, etc. The availability of temporal information in the input allow algorithms to exploit temporal consistency and context, and thus improve detection performance. However, till date, such algorithms are also evaluated using mAP, by treating each consecutive video frame as independent. Such an evaluation focuses on an object detector being able to detect a certain object in every frame. The equal weightage for every frame irrespective of the objects would often result in a negligible change in AP when certain objects are missed. For example, a detector that misses an occasional bicycle crossing the road could have a high AP since most of the frames contain parked bicycles. This may lead to life-threatening accidents [30]. A detector which consistently misses objects with specific properties of color, contrast, light, blur, speed etc. would be rated equal compared to a detector which misses certain views randomly. As an example, the mAP metric would give lesser importance to fast moving objects. This is simply because, everything else remaining the same, a faster object is likely to remain visible in lesser number of frames than a slow moving or stationary object. Similarly mAP doesn't incentivise an algorithm to detect harder instances, such as farther objects which look visibly smaller in an image.

Apart from being able to use video context to detect more objects, object detectors should also be able to use the temporal continuity to reduce false positives (Sec. 5). In a high FPS (frames-per-second) video, even a low false positive rate may imply many wrong predictions in a short time interval, which may overwhelm a user or a planning algorithm. We, therefore, argue that in video based object detection, false positives should be highly penalized.

In video object detectors, trackers can also be used to forecast object positions from previous frames. Thus, a per-frame detection of an object is a role better suited for a tracker and is rightly a part of the tracking evaluation metrics (MOTA and MOTP) [27]. We propose that for evaluation of video object detection algorithms, it would be a more appropriate objective to detect all objects in the video even if some frames/views of the object are missed in certain frames intermittently. Fig. 1 shows three detectors with precision of 1. Here, we show how similar recall values might indicate different detection capabilities. The problem lies in the assumption that each frame is independent, and detecting or missing an object in each frame carries equal weight. If we were to combine the first and last five frames in sets for each of the objects, the set recall would be higher for $\mathcal{D}_1$ (1) than $\mathcal{D}_2, \mathcal{D}_3$ (0.5).

Thus, to capture the detectability of objects throughout a video sequence, without rewarding detection of every instance, we combine the bounding boxes for an object across frames into sets. We combine bounding boxes using a unifying criteria, $\mathcal{U}$ (more details in Sec. 3.1). We argue that it is sufficient to detect an object in this unified set rather than detecting the object at every frame. For example, the frames where an object is around it's initial location, and hence look visibly similar, could be clubbed together in a set. The recall for a video is summed over all such sets within the video. The True Positives (TP*), and False Negatives (FN*) are accumulated over sets and then used to calculate the mAP, hereafter called the VmAP (Video mean average precision). Our metric can be used along with complementary metrics like Average Delay[18] to capture the delay with which the objects are detected. Our code is available at https://github.com/vaibhavg152/VmAP-video-object-detection-evaluation-metric.

Contributions: The key contributions of this paper are: **(1)** We propose a new evaluation metric (VmAP) for video object detection which is sensitive to biases in object detectors. **(2)** We demonstrate empirically that detectors ranked by Set Recall provide better post–tracking performance than Frame Recall. Ranking by set recall has shown a 9% increase in the spearman correlation coefficient with respect to the post-tracking ranks with 13 detectors. **(3)** Our metric helps choose a better operating point for video based object detectors with (upto 99%) lesser false positives leading to improved utility of these algorithms in real life applications.

2 RELATED WORK

Metrics for object detection: Object Detection is the task of predicting a rectangular (or cuboidal in case of 3D) bounding box and class scores corresponding to each object in the image. The PASCAL VOC Challenge [11] was a popular object detection contest with interpolated average precision [28] as the metric for evaluation. In 2012 [8], this was later changed to average precision for a finer comparison among methods. The predicted bounding boxes are classified as True Positives and False Positives, based on the intersection over union (IoU) overlap score with the ground truth. Any remaining objects in the scene are classified as False Negatives. These numbers are then accumulated across all frames in the data under evaluation. The Average Precision (AP) then captures the performance of the detector under varying recall values. A mean over all classes of objects produces the final score of mean average precision (mAP). COCO mAP [15], in addition, uses an average over different IoU thresholds used for matching. A detailed comparison has been done in recent surveys [20, 21, 33]. Localization-Recall-Precision (LRP) [19] re-defines the error and provides an optimization strategy for determining the best operating point for a given detector. However, this lacks the robustness of comparison over different operating points.

Related/Complementary Metrics: Mao et al. [17, 18] have proposed a metric for rewarding early detection of objects in the videos. The authors define *algorithmic delay* as the delay (in number of frames) in detecting the object once it has appeared in the video. The authors suggest finding an average of the delay at different false alarm rates for incorporating different operating conditions. While the metric addresses the issue of early detection, the evaluation still rewards detection at every frame thus biasing the evaluation towards objects with more frames. As pointed out in the paper, frame based detectors show a better average delay than aggregation based methods[34]. This shows that average delay alone doesn't indicate the superiority of a detector. From a fairness perspective also, though the average delay of unfair detector may be larger, it may be large even for the unbiased detector. This is also evident in Fig. 1, where average delay of $\mathcal{D}_1$, and $\mathcal{D}_2$ could be same, and the

metric cannot distinguish between biased and unbiased detectors. Sobti et.al. [29] propose a metric which does not reward detection at every frame. The metric targets evaluation of object detectors for real-time streams under various resource constraints[1]. Their metric, however, does not address false positives and is only applicable to videos with limited vertical motion. The metric also ignores the re-detection required for an object once its appearance/location has changed significantly. Li et al. [13] evaluate the performance of the complete pipeline of a detector, tracker and forecaster, however, it suffers from the same problems as mAP as far as sensitivity to bias is concerned. The problem of streaming perception is also different from evaluation of object detectors. Streaming Perception ranks systems in the order of state estimation quality while evaluation of object detection in videos ranks object detectors in the ability of detecting different object instances.

Tracking vs Video Object Detection: As compared to the object tracking literature[27], where the focus is on being able to track the object in every frame, we propose that the evaluation of object detectors on videos should be different. Instead of rewarding a detector that does the task of a tracker as well, the detector should be rewarded for identifying unique objects, thus generating sufficient observations (for all objects without any bias) for a tracker to work efficiently.

Bias Identification: For evaluation of object detection on videos, we propose an alternate definition for True Positives, False Positives, and False Negatives as described in Section 3. These numbers are not accumulated on all frames, but rather on all sets of an object according to a unifying criteria defined in Section 3.1. To the best of our knowledge, this is the first work which addresses video object detection evaluation while incorporating *fairness* to under-represented objects.

3 PROPOSED METRIC

An ideal metric for evaluating object detection in videos should be able to capture if a detector detects all instances of an object which are sufficiently different from each other. Consider a case of static camera capturing a static object. There is no utility in evaluating on all the frames of this video, since a detector which detects object in one frame will be able to do in the rest as well. The argument implies that the evaluation should be adaptive to the object's motion. If an object is moving quickly, the object must be detected very frequently. On the other hand, if an object is stationary, it is sufficient to detect the object at larger intervals. In this paper, we propose to incorporate it in a principled way using a new metric. To accomplish this objective, we first suggest a method to combine different bounding boxes *for the same object* in consecutive frames into a set of bounding boxes. We then change the objective function to optimize detection of more sets rather than more frames.

[1]Our metric may also address the comparison of object detectors running at different speeds (and hence under different resource constraints), however the same is not the focus of this paper, and hence not experimentally validated.

3.1 Set Formation

In Fig. 1, say, the first 5 frames are combined into one set and the last 5 into another. We can see that, detector $\mathcal{D}_1$ detects all the sets, $\mathcal{D}_2$ detects sets only for the pedestrian B and $\mathcal{D}_3$ detects only one set per pedestrian. Thus, if we order the detectors on the basis of their performance at the set level, then this would better reflect the performance achievable through a tracker or as would be desired by a control system which uses the detections for path planning/subsequent actions. The following sections describe the criteria for forming sets and scoring the object detectors based on the objective of detection in each set instead of each frame.

Unifying Criteria: There can be different ways to unify the object instances into sets. In this section, we discuss a location-based unifying criteria $\mathcal{U}_l$. For two bounding boxes b and b' with coordinates $<x, y, w, h>$, and $<x', y', w', h'>$, the unifying criterion $\mathcal{U}_l$ is defined as:

$$\mathcal{U}_l(b, b') = \max_{\Delta x=-\gamma, \Delta y=-\gamma}^{\Delta x=+\gamma, \Delta y=+\gamma} \text{IOU}(<x, y, w, h>, <x' + \Delta x, y' + \Delta y, w', h'>) \quad (1)$$

The criteria $\mathcal{U}_l$ allows us to define a broader area around the original box where a non-zero overlap would be possible. This allows the object to move by some amount before the IOU value starts to decrease. We select γ as 10 pixels in our experiments. We use an absolute number of pixels instead of a value that is proportional to the size of the original box, so that the area is relaxed to the same extent for both small and large area bounding boxes. This implies that a similar amount of displacement would be required for both a small and a large object to form a new set. This creates an implicit reward for the object detectors to focus equally on different objects. Fig. 2 shows examples of the sets formed using our criteria.

Set Formation Procedure: For N objects in a video, each object $\{O_i\}_{i=1}^{N}$ is present in multiple frames. Let i be the index over objects, j be the index over sets and k be the index over frames. Let b_i^k represent the bounding box of O_i in the k^{th} frame. Let $|O_i|$ denote the number of frames for which the object O_i is present in the video. The set $S(i, j)$ represents the j^{th} set for the object O_i. Each set member is a tuple $<k, b_i^k>$. Note that the set is *well-ordered* with respect to k. We use the following procedure for combining the bounding boxes into various sets:

(1) For every object O_i, the well-ordered set $S(i, 0)$ is initialized with $<f, b_i^f>$, where f is the first frame when O_i appears in the video.
(2) The element $<f+1, b_i^{f+1}>$ is appended into $S(i, 0)$ if $\mathcal{U}_l(b_i^f, b_i^{f+1}) > 0$. Otherwise, a new set $S(i, 1)$ is created.
(3) For every new frame, the bounding box b_i^k is compared to the bounding box of the minimal element of the current set $S(i, j)$ (extracted using $b_{\min}(S(i, j))$) where $b_{min}(.)$ returns the minimal element in the set. If $\mathcal{U}_l(b_i^k, b_{\min}(S(i, j))) > 0$), then $<k, b_i^k>$ is appended into the existing set $S(i, j)$.
(4) Otherwise, a new set $S(i, j + 1)$ is instantiated with $<k, b_i^k>$ as its first element, and Step 3 is repeated with $S(i, j + 1)$ as the current set.

Fig. 3 demonstrates the set creation process with an example. With the method as described above, all the bounding boxes within each

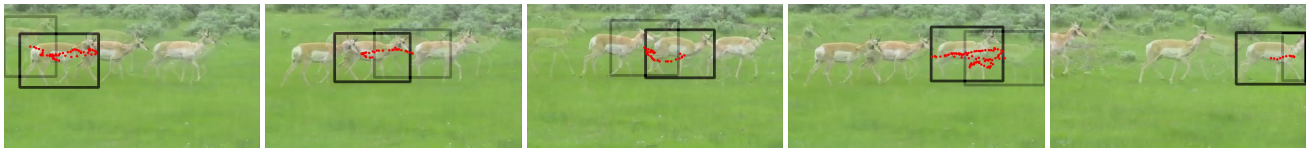


Figure 2: In this figure, we show the start (dark rectangle) and end (light rectangle) frame for three objects in video. We also show the centre of bounding box in intermediate frames (red dots). The sets contain bounding boxes of an object in a similar spatial neighborhood (as decided by the unifying criteria), as its initial location in the consecutive video frames. We blend the frames in each set for visualization.

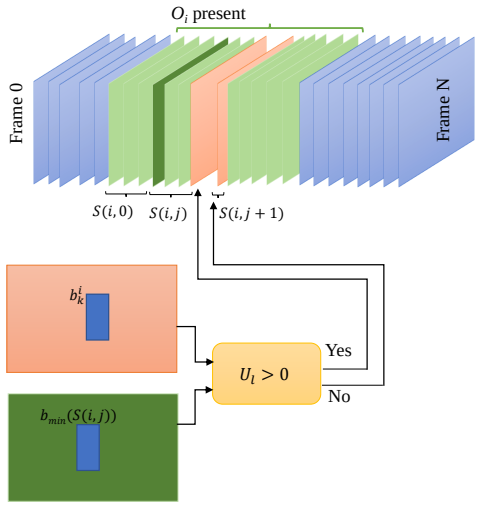


Figure 3: For every object O_i, the bounding boxes in the frames which contain the object i are combined into sets. The object bounding box in every new frame (orange frame) is compared against the minimal (first frame) of the last set (green frame from $b_{min}(S(i,j))$). If the location constraint $\mathcal{U}_l$ is above a threshold, the new bounding box is added to the previous set $S(i,j)$. Otherwise, a new set $S(i,j+1)$ is instantiated.

set $S(i,j)$ satisfy the following constraint:

$$\mathcal{U}_l(b_{min}(S(i,j)), b_k^i) > 0 \; k \; \forall \; <k, b_k^i> \in S(i,j) \tag{2}$$

Note that the procedure described above leads to different sets for each object. Since the object *id* is used from the ground truth for forming the sets, there is no incorrect assignment in case of occlusion or crowds.

3.2 Scoring

After the sets are formed, we redefine the True Positives and False Negatives at the set-level for the purpose of evaluation. Thus, both the Precision and Recall are re-defined.

True Positives (TP*): The detections are now evaluated at the set level. Consider a set $S(i,j)$, i.e., the j^{th} set for the i^{th} object, which contains bounding boxes from the frames k_1–k_2. Note that $S(i,j)$ would have similar views of the object O_i at approximately the same spatial location in different frames. If any of the detection corresponding to the frames k_1 to k_2 match the ground truth for the respective frame, the set is counted as a true positive. This choice of *any one* detection in a set is further explored in the supplementary material. A detection d_l is said to match the ground truth bounding box b_i^l, if the bounding box d_l has an intersection over union score IOU > $\text{IOU}_{\text{thresh}}$ with the bounding box b_i^l in frame l within the set $S(i,j)$:

$$\text{IOU}(d_l, b_i^l) > IOU_{\text{thresh}} \tag{3}$$

The matching is done in the decreasing order of confidence level for the object class as originally discussed in [9].

False Negatives (FN*): Any set $S(i,j)$ which has no matches d_l is regarded as a False Negative. This indicates that the detector missed an entire range of appearances/locations of that particular object.

False Positives (FP*): False positives in a video are particularly important. A false positive signifies that an object has been detected (possibly by confirmation from multiple frames). Therefore, in the proposed metric each false positive is counted without assignment to any particular set. This is why it becomes difficult to obtain high precision using our video metric, since any singular wrong frame based false positive will lead to a false positive in our metric. This is in contrast to computation of false negative where a single frame based false negative does not necessarily lead to a false negative in our metric, as long as the detection in all the frames in a set are not missed. We believe that asymmetric counting of false positives/negatives is reasonable in video based object detector, since low false positive rate is particularly important for the video object detection due to their potential to cause repeated and frequent disruptions in an high FPS video. Note that since all false positives are accumulated across frames, it may make sense for an algorithm to intentionally skip reporting detections which have been found only on few frames while processing a video to ensure that only high confidence detections are reported. The video-level precision (VP) and recall (VR) is then defined as:

$$\text{VP} = \frac{\text{TP}^*}{\text{TP}^* + \text{FP}^*}, \quad \text{VR} = \frac{\text{TP}^*}{\text{TP}^* + \text{FN}^*} \tag{4}$$

Similar to Everingam et al. [10], we use the average precision at different recall values and a mean across different classes to get the VmAP (Video Mean Average Precision).

The ablation studies for the set formation criteria and the matching procedure are discussed in the supplementary material.

4 EXPERIMENTS

In this section, we discuss results of the experiments corresponding to two primary claims for our metric:

(1) Our metric brings out the biases in object detectors. If a detector has a certain bias against some objects, our metric would score it significantly lower while the mAP metric scores it at par with other detectors.
(2) Using the Video Recall and Video Precision curve, we obtain a much better operating point. The operating point has much lesser false positives while having a similar number of sets detected.
(3) The ranking of object detectors as determined using Video Recall (VR) has a higher correlation with the post-tracker performance of the detectors as compared to frame recall.

Dataset Details: To compare object detection algorithms in videos, we use the validation set from the Imagenet VID dataset [26]. The VID validation set contains 30 classes of objects in 555 snippets. The annotations for the dataset contain object *id* as well as the bounding box location in each frame.

Object Detectors: We use object detectors trained on the VID dataset, namely, RDN [6], FGFA [34], MEGA [4] and DFF [35]. These detectors take advantage of multiple frames both global as well as local aggregation before predicting the bounding boxes per frame. For a wider analysis, we also test the efficacy of models trained on the COCO dataset [15] on 347 snippets from the VID dataset having common classes of objects between VID and COCO datasets. The detectors in this category are Centripetal Net [7], Corner Net [12], Faster RCNN [25], YOLOv3 [24], DETR [1], FCOS [31], HTC [2], and, Retina Net [14]. These detectors include object detectors using transformers, keypoint regression, center regression, Feature Pyramid Networks (FPN) as well as two stage detectors. We would like to acknowledge MMDetection Library [3] for providing pre-trained models for the same.

4.1 Fairness against Biases

Evaluating video object detectors using frame based evaluation metrics fails to capture and penalize any bias that might be present in the detector. In this section, we experimentally show this insensitivity of mAP towards biases in a detector. For bringing out this effect, we first create synthetic detectors (Sec. 4.1.1) by artificially adding biases against some aspects in the detection. Finally, we show how the mAP can be manipulated (Sec. 4.1.2) by varying the number of frames/objects in a video. The mAP metric is unable to capture the bias of detectors on the real dataset.

4.1.1 Synthetic Detectors on Real Data. As a first step to create synthetic detectors with various biases, we add new bounding boxes at random locations and size, and equal to the number of ground truth bounding boxes. The confidence values for the bounding boxes (both ground truth and random) are sampled from a log-normal probability distribution. For creating an *unbiased* detector, we fix $p = 0.5$ as the probability that each bounding box will be outputted by the detector. Thus, all objects are equally likely to be detected.

We introduce bias in the detector by tweaking the probability of a box being outputted by the detector. The probability is tweaked for ground truth boxes with certain attributes to create a bias. Further, to make the biased and unbiased detectors directly comparable, we keep the total number of detections similar in the two detectors. We achieve this by decreasing the probability of outputting rest of the boxes. We explain below with an example.

Consider an unbiased detector which was generating 500 detections, of which 50 boxes were of size more than 1000. Let us consider a biased detector which favors large objects. To create such a synthetic detector a larger ground truth bounding box with area more than 1000 may have its probability of selection by the biased detector set to $p = 0.75$. Now let us assume that after increasing p for larger objects, the total number of output boxes increases to 600, of which 150 are large objects. For reasons as described above, we also reduce the probability of rest of the boxes in the ground truth such that their total number in the output comes to 350, and the total number of detections remain at 500.

The above method of setting probabilities allows us to create plausible detectors with different biases but a similar mAP value. This implies that on the basis of mAP these detectors are not distinguishable. We then evaluate such detectors using proposed VmAP to see if our metric can expose their biased nature and score them lower.

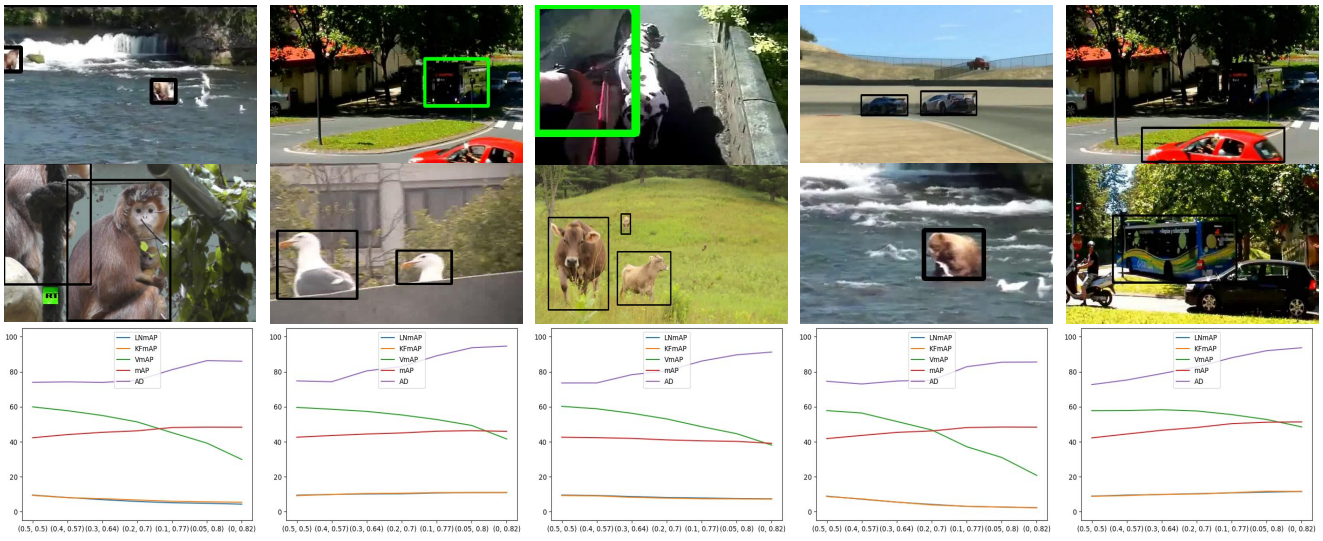
For the experiments below, to create bias against particular type of objects, their probability p is decreased from 0.5 to 0.0 (in steps), and for the remaining boxes, p is increased in a manner described above. We introduce following biases in the synthetic detectors:

- *Size:* A bounding box is declared small if its area (in pixels squared) is less than 4% of the total area of the image. The detector is biased against small objects.
- *Brightness:* An object is termed dark if the average luminance of pixels within the bounding box is < 90. Dark boxes are biased against.
- *Contrast:* An object having a contrast < 0.42 is made less likely to be detected. To quantify the color contrast, we use Weber Contrast[22]:

$$\text{Contrast} = \frac{I - I_b}{I_b} \tag{5}$$

  where I represents the average luminance of the pixels in the bounding box and I_b represents that of the background. For a bounding box of dimensions $l \times b$, a region of $1.5l \times 1.5b$ around the bounding box (at the same center) was used as the background.
- *Speed:* A fast moving object is defined as the object having a set size less than 40 frames. The probability of faster objects getting detected is reduced.
- *Color:* The synthetic detector is made less likely to detect objects having a red color. Hue of the dominant color inside the bounding box is calculated and if it lies within [-60, 60] (red range), probability of it being detected is reduced.

Fig. 4 shows the comparison plots for the biased detector described above. The x-axis shows various degrees of bias in these detectors, whereas y-axis shows various metric scores. We note that while the VmAP decreases with increase in bias, the mAP remains constant. This confirms that the proposed metric can be used to detect, and lowers the score of biased detectors.



(a) Size (b) Illumination (c) Contrast (d) Speed (e) Color

Figure 4: The figure shows examples of the biases introduced and behavior of mAP and VmAP on increasing these biases. First row shows the examples of objects less likely to be detected (small/dark/low contrast/fast moving/color) while the second row shows examples of their counterparts which are more likely to be detected. The X axis shows the increasing degree of biases and Y axis shows the behaviour of mAP and VmAP. mAP is fairly insensitive to biases in all the cases, whereas VmAP reduces on increasing the bias. The amount of sensitivity of VmAP varies in each case. Other metrics like mAP normalized with length of set (LNmAP) and mAP of just keyframes (KFmAP) are also found insensitive. Average Delay[18] also increases with increase in bias of the detectors. The metrics are defined in detail in the supplementary material.

To test if alternate formulations of metrics are able to capture this bias, we also include Average Delay Metric [18] and two more baselines - LNmAP (Length-Normalized mAP) and KFmAP (mAP of the keyframes). The length-normalized mAP normalizes the number of frame-wise true/false positives by the length of the set. The KFmAP picks every 10^{th} frame(as used in [35]) as the keyframe and finds the mAP using only these keyframes. KFmAP and LNmAP follow the mAP curve with the variation in bias, thus unable to capture the bias in the detectors.

4.1.2 Dataset Curation to Trick mAP Evaluation. To show how the mAP metric is susceptible to manipulation using the dataset, we curate some toy examples from the Imagenet VID dataset[26]. In Tab. 1, in first three rows, we increase the number of small objects ($D_1 - D_3$) while keeping the overall number of frames corresponding to small objects same. In curated dataset D_1, we start with 600 frames from two videos - one containing a large bear and another containing a small bear. In dataset D_2, we replace half the frames with frames from a different video, resulting in similar number of frames for small objects but increasing the objects from 1 to 2. Similarly in D_3 we increase the number of objects further to 4, while keeping the frames corresponding to them to 600. We use these 3 datasets to evaluate a detector (RDN [6]) which we believe is biased against small objects. However, note that the mAP values remains same for the detector on the three datasets. On the other hand, VmAP is able to successfully degrade the detector.

By careful manipulation of the dataset, a biased detector may be made to look good as shown above. Similarly another choice of dataset may be made to look a detector bad as well. In Tab. 1, rows 4-6, we incrementally add the frames from a video containing small objects. We test DFF [35] on these datasets, and the mAP shows that the same detector is worsening. Whereas VmAP successfully maintains the performance of the detector at the same level.

The above experiments show that even for the real detectors, one can choose the datasets carefully to trick the mAP comparison. Our proposed metric is successful in discerning such nuances, and give expected scores to the detectors.

4.1.3 Biased Detectors: Visual examples. Fig. 5 shows some visual examples from the videos where the unfair behavior of a detector is masked by the mAP based evaluation. The timeline plots in the figure denote the presence, set formation as well as the detection of objects. The rectangles represent sets of objects. Thus, one can see that O_0 stays at around the same location for frames 0–50. The green and red dots represent frame-wise detects and misses. The rectangle is shaded green in case of a True Positive set and red in case of a False Negative set. False positives per frame are shown in the upper axis.

4.2 Operating Point

When using a detector in a practical application, often the most important metric is the accuracy at the best-suited threshold. The

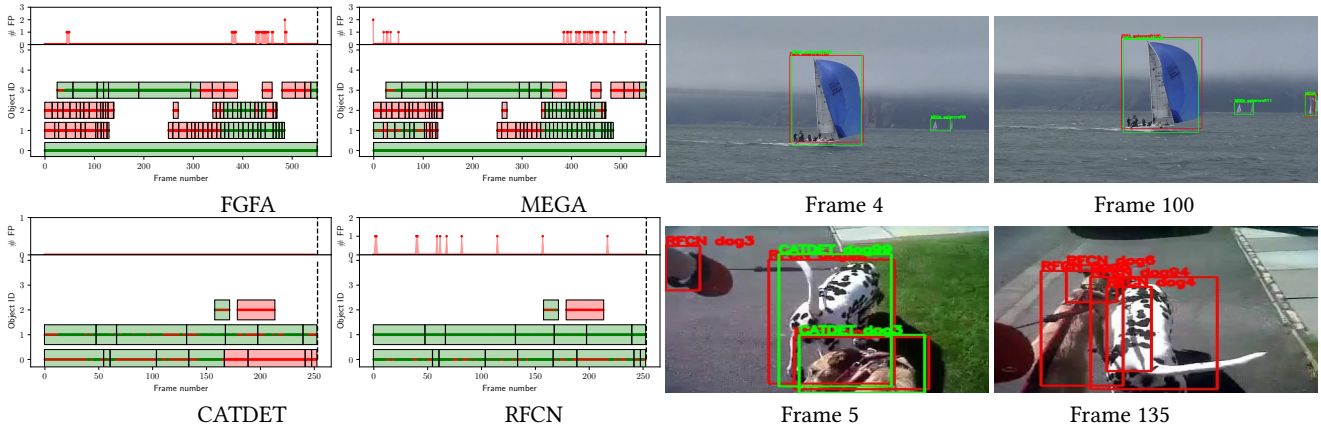


Figure 5: The timeline plots for a video (ref. Sec. 4.1.2) for two different detectors are shown on the left side. In the first row, FGFA (mAP 78.0) and MEGA (mAP 75.1) have similar mAP accuracy. However, FGFA fails to detect the small watercraft in the video. This is captured by the VmAP metric (MEGA - 57.8 Vs FGFA - 44.5). Similarly, in the second row, CATDET is able to detect almost all sets of the object without having a single false positive, while RFCN has multiple false positives during the video. The similar VmAP (CATDET - 81.8 Vs RFCN - 77.6) despite a vastly different mAP (CATDET - 23.8 Vs RFCN - 61.2) can be observed from the number of sets that are detected.

	f_s	f_l	O_s	O_l	mAP	VmAP
D_1	600	600	1	1	49.6	50.0
D_2	532	600	6	2	48.8	38.3
D_3	600	600	11	4	46.7	30.3
D_A	60	360	1	1	98.3	53.5
D_B	180	360	1	1	95.8	55.2
D_C	360	360	1	1	91.4	55.2

Table 1: The table shows mAP, and VmAP values for the same object detector (biased against small objects) when tested on different subsets of a dataset. Notice that depending upon which videos we include or exclude in the dataset, the detector can be made to look arbitrarily good or bad. f_s(f_l) and O_s(O_l) represent number of frames, and number of sets in the small and large category respectively. Refer Sec. 4.1.2 for the detailed discussion.

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
DFF	86.96	84.78	108	1
FGFA	89.13	82.61	105	1
RDN	91.3	86.96	258	3
MEGA	95.65	93.48	109	2

Table 2: The table shows the set recall (VR) and false positives (FP) with the P-R operating point and the VP-VR operating point on videos of hamster class from the Imagenet VID dataset (other classes in supplementary). It is observed that Video Precision/Recall operating point has much lesser false positives while having a similar set recall.

best-suited threshold is typically defined using the maximum F1-score point on the PR curve[5, 16]. Tab. 2 shows the number of false positives from various popular detectors when operated at the best threshold computed from VP, and VR as defined in Eq. (4). We compare it with the false positives obtained from the best threshold computed from frame level precision and recall. Note that while the VR remains similar at both the threshold values, the number of false positives decrease from 108 to 1 for DFF [35], and from 258 to 3 for RDN [6].

Thus, detectors deemed better at the frame level may not perform better in a video based system. Our experiments shown earlier indicate that VmAP could be a better proxy for performance at the system level.

4.3 Post-Tracking Assessment

Detectors are typically ranked in isolation from other components of the system, like trackers, or even planning and control systems. However, in practical systems, the detectors are often run with trackers to increase the confidence in the predicted objects, as well as, to establish an association between the predicted boxes in different frames. Consider an object detector being run at a certain frame precision, say 0.9. The efficacy of the detector is typically measured using the Frame Recall. However, we argue that the "ground truth ranking" for the detectors is better indicated by the performance of the object detector with an ideal tracker. An ideal tracker predicts the bounding box location of a detected object in the next frame. It also filters out false positives in frames where sufficient evidence of an object detection is not found. Thus, we use the Frame Recall of an object detector coupled with an ideal tracker with a Frame Precision $P = 0.9$. The order of frame recall is considered the ground truth ranking for the object detectors. In Tab. 3 we present the rankings of the detectors with Frame Recall (FR) and Video Recall (VR) as the criteria as well as the ground truth rankings (FR post-tracking). VR has a spearman correlation coefficient (SCC) of 0.97 while FR has an SCC of 0.93 on the entire dataset. We further verified that on a 10-way split of the dataset, VR

Detector	FR	Rank	VR	Rank	FR*	Rank
RDN	80.45	2	79.27	1	**92.27**	**1**
MEGA-BASE	77.85	4	75.95	3	**91.92**	**2**
MEGA	81.6	1	77.12	2	**91.31**	**3**
YOLOV3	73.18	6	71.52	5	**90.75**	**4**
FGFA	79.98	3	72.75	4	**89.89**	**5**
DFF	75.57	5	69.93	6	**88.33**	**6**
DETR	56.37	9	65.74	8	**85.33**	**7**
FCOS	60.58	7	66.44	7	**84.93**	**8**
HTC	58.99	8	63.01	9	**82.67**	**9**
RetinaNet	56.18	10	60.45	11	**82.02**	**10**
Centripetal	53.24	11	62.31	10	**80.79**	**11**
FRCNN	49.69	12	56.65	12	**77.17**	**12**
CORNERNET	47.04	13	55.48	13	**76.48**	**13**

Table 3: We test the ranks of detectors as determined by the Frame Recall (FR) and Video Recall (VR) at 0.9 frame precision. The ground truth order is determined by using the Frame Recall (FR*) when an ideal tracker is used along with the detector at the given precision. Video Recall (VR) achieves a spearman correlation of 0.97 as compared to 0.93 for Frame Recall (FR) indicating the suitability of VR as a measurement for post-tracking effectiveness.

and FR have an SCC of 0.95 ± 0.013 and 0.87 ± 0.023 respectively. The corresponding rank error is 8 and 14. This further highlights the effectiveness of the set formation strategy and evaluation in sets. The trend of VR, FR and FR with ideal tracker for the detectors is also shown in the supplementary material.

4.4 Ablation Studies

We present various ablation studies to understand the effect of different hyper-parameters in our system in the supplementary material. We analyze various unifying criteria to combine bounding boxes into sets. For example, we have considered appearance based unification as well as location based, and their various combinations. We also do an ablation study by varying the definition of a true positive set.

5 DISCUSSION

The progress in object detection methods for individual images was marked by the addition of datasets. There is now significant amount of data available for video object detection as well[23, 26]. The recent trend in video object detection has been to use more and more context locally (from nearby frames) as well as globally (from far-away frames)[4] due to which the mAP numbers have increased. Local context enables the detectors to ensure that the detection effort from the previous frame is not ignored when detecting the next frame. While many object detectors [4, 32, 34] take advantage of other frames to get object detections, they do not explicitly reduce false positives using this mechanism. As we take focus away from per-frame detection using our metric, we also propose to take the focus towards robust (very few false positives) and fair (all types of objects) object detection in videos.

In Fig. 6, we show how simply accumulating detections over 3 frames before reporting the detections increases the precision at the operating point by upto 100%. The bounding boxes from 3 frames

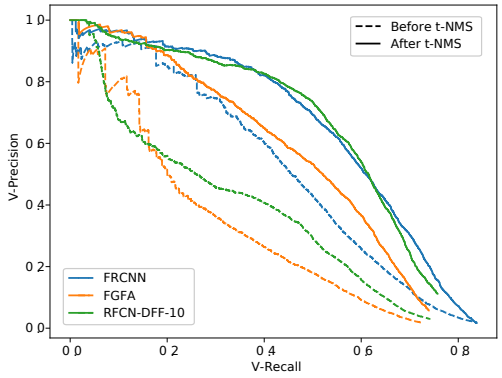


Figure 6: The PR curve shows the effect of temporal NMS across frames. Higher recalls now have better precision due to reduced false positives.

are accumulated on a single frame and a (temporal) non-maximal suppression is done on the resultant boxes. This removes any false positives in the vicinity of the detected object and allows a detector to pick the best detection from 3 frames. With this simple strategy for false positive reduction, we suggest that there is a scope in video object detectors to look closely at false positives and develop strategies for more confident detections of objects.

6 CONCLUSION

The proposed metric, Video Mean Average Precision (VmAP), reveals alternate goals for progress in video object detection methods where the focus shifts from per-frame detection to detecting objects more confidently and without omission of any objects. Our metric fairly evaluates objects moving at different speeds in the video sequence, thus equally weighing objects present at the same location for hundreds of frames and a fast moving object present at a location for just 2-3 frames. The results of our evaluation expose weakness of the current evaluation metric in controlling the false positive rates at a video level and an unhealthy focus on per-frame detection. The strategies for false positive reduction show a path forward towards further improvement of video object detection methods in a way that would be useful in a complete system, be it an autonomous car making decisions about how to advance or a visually impaired person listening to description of objected detected.

ACKNOWLEDGMENTS

Anupam Sobti was supported by Visvesvaraya PhD Scheme, MeitY, Govt of India (MEITY-PHD-1292). This work has been partly supported by the fundings received from DST through the IMPRINT program (IMP/2019/000250) and ICPS IoT Cluster 2018.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and

Dahua Lin. 2019. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[4] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory Enhanced Global-Local Aggregation for Video Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10337–10346.

[5] Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. 2011. An exact algorithm for F-measure maximization. *Advances in neural information processing systems* 24 (2011), 1404–1412.

[6] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. 2019. Relation distillation networks for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 7023–7032.

[7] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. 2020. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10519–10528.

[8] M Everingham, L Van Gool, CKI Williams, J Winn, and A Zisserman. 2012. The pascal visual object classes challenge 2012 results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results. (2007).

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop.

[12] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *15th European Conference on Computer Vision, ECCV 2018*. Springer Verlag, 765–781.

[13] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. 2020. Towards Streaming Perception. In *European Conference on Computer Vision*. Springer, 473–488.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[16] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize F1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 225–239.

[17] Huizi Mao, Taeyoung Kong, and William J Dally. 2018. Catdet: Cascaded tracked detector for efficient object detection from video. *arXiv preprint arXiv:1810.00434* (2018).

[18] Huizi Mao, Xiaodong Yang, and William J Dally. 2019. A Delay Metric for Video Object Detection: What Average Precision Fails to Tell. In *Proceedings of the IEEE International Conference on Computer Vision*. 573–582.

[19] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2018. Localization recall precision (LRP): A new performance metric for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 504–519.

[20] R. Padilla, S. L. Netto, and E. A. B. da Silva. 2020. A Survey on Performance Metrics for Object-Detection Algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 237–242.

[21] Rafael Padilla, Wesley L Passos, Thadeu LB Dias, Sergio L Netto, and Eduardo AB da Silva. 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics* 10, 3 (2021), 279.

[22] Eli Peli. 1990. Contrast in complex images. *JOSA A* 7, 10 (1990), 2032–2040.

[23] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. 2017. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5296–5305.

[24] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[27] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *arXiv preprint arXiv:1701.01909* (2017).

[28] G SALTON and MJ MCGILL. 1986. Introduction to Modern Information Retrieval (pp. paginas 400).

[29] Anupam Sobti, Chetan Arora, and M Balakrishnan. 2018. Object detection in real-time systems: Going beyond precision. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1020–1028.

[30] Neville A Stanton, Paul M Salmon, Guy H Walker, and Maggie Stanton. 2019. Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. *Safety Science* 120 (2019), 117–128.

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9627–9636.

[32] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence Level Semantics Aggregation for Video Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9217–9225.

[33] Haidi Zhu, Haoran Wei, Baoqing Li, Xiaobing Yuan, and Nasser Kehtarnavaz. 2020. A Review of Video Object Detection: Datasets, Metrics and Methods. *Applied Sciences* 10, 21 (2020), 7834.

[34] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 408–417.

[35] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2349–2358.

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
horse				
MEGA-BASE	54.5	35.98	467.0	50.0
DFF	45.5	26.98	441.0	1.0
FGFA	46.56	32.28	656.0	93.0
RDN	54.5	33.86	452.0	38.0
MEGA	55.03	37.57	408.0	38.0
lion				
MEGA-BASE	73.33	60.0	5.0	0.0
DFF	66.67	66.67	1.0	0.0
FGFA	60.0	60.0	5.0	0.0
RDN	100.0	100.0	26.0	5.0
MEGA	100.0	100.0	22.0	0.0
car				
MEGA-BASE	82.85	45.65	5590.0	207.0
DFF	77.7	38.92	6565.0	208.0
FGFA	80.74	44.2	6028.0	260.0
RDN	84.43	52.64	5741.0	317.0
MEGA	81.53	51.19	5554.0	373.0
cattle				
MEGA-BASE	73.06	57.51	233.0	14.0
DFF	72.02	55.44	437.0	39.0
FGFA	62.69	53.37	345.0	23.0
RDN	78.76	63.73	515.0	31.0
MEGA	83.94	67.36	553.0	34.0
airplane				
MEGA-BASE	84.42	77.39	387.0	78.0
DFF	76.88	69.72	573.0	83.0
FGFA	80.53	68.59	615.0	57.0
RDN	86.06	78.89	445.0	74.0
MEGA	85.55	77.64	403.0	91.0
antelope				
MEGA-BASE	91.34	87.4	122.0	15.0
DFF	88.19	78.74	125.0	15.0
FGFA	88.98	81.89	107.0	30.0
RDN	92.91	90.55	96.0	40.0
MEGA	71.65	71.65	26.0	26.0

Table 5: Table showing false positive comparison of PR and VP-VR operating points

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
bear				
MEGA-BASE	91.6	57.14	633.0	31.0
DFF	87.39	67.23	561.0	22.0
FGFA	94.12	82.35	666.0	326.0
RDN	95.8	45.38	892.0	26.0
MEGA	94.12	56.3	775.0	36.0
bicycle				
MEGA-BASE	78.88	57.37	525.0	65.0
DFF	68.92	53.39	576.0	77.0
FGFA	70.52	53.78	399.0	42.0
RDN	77.29	58.96	542.0	56.0
MEGA	74.5	54.18	452.0	27.0
bird				
MEGA-BASE	41.46	28.85	383.0	28.0
DFF	35.29	35.01	132.0	96.0
FGFA	33.89	30.81	189.0	60.0
RDN	48.74	36.69	558.0	94.0
MEGA	41.46	36.41	210.0	100.0

Table 6: Table showing false positive comparison of PR and VP-VR operating points

Detector	mAP	$\mathcal{U}_l$	$\mathcal{U}_a$	$\mathcal{U}_{al}$	$\mathcal{U}_t$
MEGA	83.4	56.6	71.6	66.1	76
RDN	81.3	53.65	67.3	60.6	72.5
FGFA	78.9	49.4	63.6	55.4	69.3
DFF	75.3	45.6	59.9	51.3	65.2

Table 4: The table shows that although the range of values are quite different, there isn't any difference in the ranking of algorithms by VmAP with different set formation strategies.

A ABLATION: UNIFYING CRITERIA

We experiment with the following unifying criteria:

- **Location:** The criterion described in the main paper and used in other experiments.
- **Appearance:** We use 3D histogram of RGB values of pixels inside the bounding box b and use the resulting 512-dimensional vector (${V^b}_{app}$) as the feature representing the appearance of b. For bounding boxes b_1 and b_2, the unifying criteria $\mathcal{U}_a$ is defined as:

$$\mathcal{U}_a(b_1, b_2) = ||{V^{b_1}}_{app} - {V^{b_2}}_{app}|| \tag{6}$$

  where $||.||$ represents the euclidean distance between vectors The bounding box b_2 is added to the set if $\mathcal{U}_a(b_1, b_2) < \theta_{app}$, where θ_{app} is the threshold.
- **Location + Appearance:** A bounding box b_2 is added to the set with b_1 as its first frame if $\mathcal{U}_a(b_1, b_2) < \theta_{app}$ and $\mathcal{U}_l(b_1, b_2) < \theta_{IoU}$
- **Time duration:** A bounding box is added to a set if the number of frames already present in the set is less than a certain threshold (θ_t).

The results are shown in Table 4. There is not much difference in the ranking of the algorithms while the absolute numbers change as per the criteria. Fig. 7 shows an example of sets formed in the same video with different unifying criteria. As one may observe,

Figure 7: The figure shows the first and last frame of two of the sets formed by using the location criteria (left) and appearance criteria (right). In the upper row, the appearance criteria declares a new set when the object has come too close. This might be dangerous for perceive and control systems like self driving cars. In the second row, the appearance criteria declares a new set even when the train hasn't moved much. In both the cases, the location criteria gives a better estimation.

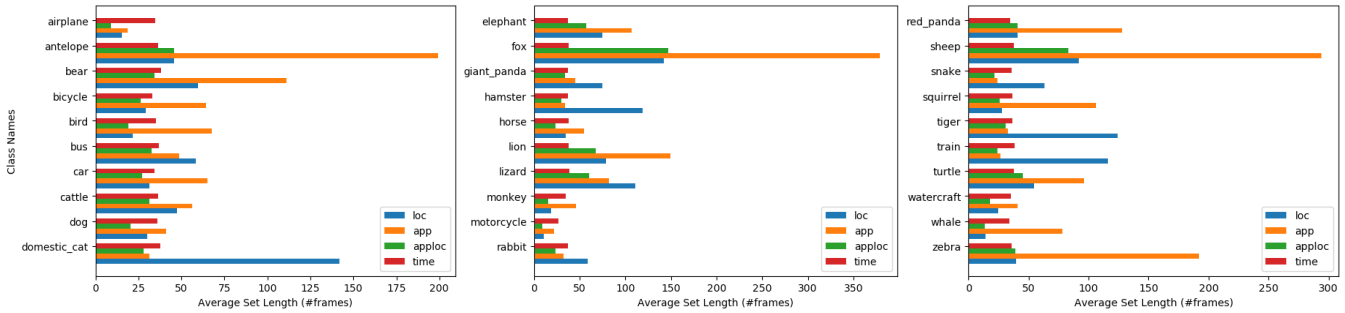


Figure 8: The sets are smaller for the appearance ($\mathcal{U}_a$) and even smaller for the ($\mathcal{U}_{al}$) criteria. As expected, the set size for the time criteria is the same ($\mathcal{U}_t$). We believe that all criteria could be useful in different scenarios as discussed in Sec. A.

the appearance criteria may be more suited in counting like applications, while the location criteria is more suitable to perceive and control systems (for optimal path planning and control tasks, etc.).

Figure 8 shows the average set length for each class. The sets are smaller for the apperance and appearance-location criteria. As expected, the set size is fixed for the time criteria. We believe that all criteria are useful in some scenarios, e.g., location criterion in perceive-and-control systems, the time criterion in a latency-critical application and the apperance-location criterion for safety-critical applications where every small change needs to be tracked.

B ABLATION: VMAP DEFINITION

In the main paper, we define a True Positive Set as a set in which at least one of the bounding boxes of the object in the set has an $IOU > IOU_{th}$. One may question the decision of taking at least one of the bounding boxes in the set. In this section, we verify whether the number of frames detected within a set is a critical parameter. We do this by performing the experiments with synthetic detectors as done in the main paper. The following definitions are compared:

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
bus				
MEGA-BASE	67.06	56.47	398.0	44.0
DFF	64.71	43.53	321.0	7.0
FGFA	57.65	50.59	156.0	28.0
RDN	69.41	52.94	292.0	12.0
MEGA	72.94	55.29	508.0	26.0
dog				
MEGA-BASE	79.81	57.08	1220.0	122.0
DFF	71.46	52.9	1489.0	120.0
FGFA	79.35	57.31	929.0	71.0
RDN	83.06	62.41	1041.0	131.0
MEGA	85.15	65.66	1078.0	96.0
domestic cat				
MEGA-BASE	77.27	27.27	307.0	0.0
DFF	63.64	27.27	437.0	41.0
FGFA	81.82	22.73	242.0	5.0
RDN	72.73	36.36	188.0	0.0
MEGA	13.64	9.09	33.0	0.0
elephant				
MEGA-BASE	90.09	61.26	749.0	8.0
DFF	89.19	61.26	502.0	40.0
FGFA	85.59	56.76	556.0	21.0
RDN	92.79	73.87	772.0	30.0
MEGA	95.5	81.98	1019.0	27.0
fox				
MEGA-BASE	61.9	57.14	15.0	2.0
DFF	66.67	57.14	46.0	1.0
FGFA	66.67	61.9	30.0	3.0
RDN	71.43	66.67	28.0	3.0
MEGA	76.19	66.67	28.0	3.0
giant panda				
MEGA-BASE	70.97	59.68	164.0	9.0
DFF	70.97	53.23	373.0	2.0
FGFA	69.35	59.68	188.0	1.0
RDN	77.42	53.23	228.0	1.0
MEGA	30.65	30.65	3.0	3.0

Table 7: Table showing false positive comparison of PR and VP-VR operating points

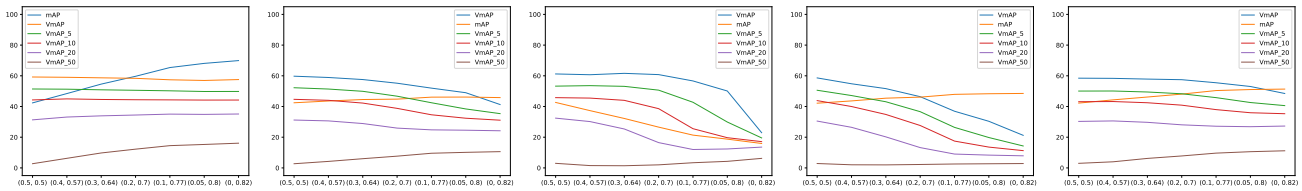
(1) **VmAP** This is the baseline used in the paper. At least one bounding box in the set must have an $IOU > IOU_{th}$ as compared to the ground truth boxes.
(2) **VmAP_*N***: An alternative to using a single frame per set is to use a percentage of the set length as a threshold for counting a true positive. For example, in the VmAP_5 definition, a set is considered true positive if the detected bounding box matches the ground truth boxes in at least 5% of the frames within the set.

C OPERATING POINT COMPARISON

Tables 5, 6 and 7 show the reduction in false positives when using the confidence threshold corresponding to the optimal point on the P-R and VP-VR curves. As shown in the main paper, the number of false positives always decreases. In some cases, the decrease in set recall (VR) is also observed, which can be improved using traditional false-positive/false-negative trade-off using confidence threshold tuning.

D POST-TRACKING ASSESSMENT

As discussed in the main paper, the Video Recall (VR) measure follows the FR + Tracker measure more closely as compared to the Frame Recall (FR) measure. Fig. 10 shows this trend in the decreasing order of ground truth measure (FR + Tracker). This indicates the suitability of the metric for measurement of the true ability of an object detector, when it would be used in conjunction with a tracker.



(a) Size (b) Illumination (c) Contrast (d) Speed (e) Color

Figure 9: The figure demonstrates how the different VmAP definitions respond to increase in different types of bias. We find that as the % value on number of frames in a set is increased, the behavior of the metric resembles that of mAP. In higher percentage values, a true positive set becomes very difficult to achieve resulting in extremely low scores. We, therefore, recommend the usage of VmAP as defined in the main paper.

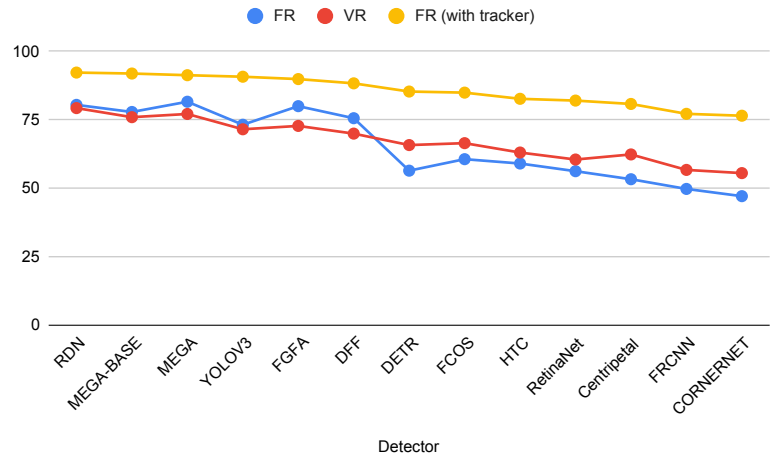


Figure 10: The detectors are presented in the order of ground truth rankings (Frame Recall using a Detector + Ideal Tracker). While the Frame Recall is low for some detectors (see FGFA, DETR), the post-tracking performance is quite high. On the other hand, VR follows the post-tracking performance closely.