

# Analysis and a Solution of Momentarily Missed Detection for Anchor-based Object Detectors

Yusuke Hosoya<sup>1</sup> Masanori Suganuma<sup>1,2</sup> Takayuki Okatani<sup>1,2</sup>

<sup>1</sup>Graduate School of Information Sciences, Tohoku University <sup>2</sup>RIKEN Center for AIP

{yhosoya, suganuma, okatani}@vision.is.tohoku.ac.jp

## Abstract

The employment of convolutional neural networks has led to significant performance improvement on the task of object detection. However, when applying existing detectors to continuous frames in a video, we often encounter momentary miss-detection of objects, that is, objects are undetected exceptionally at a few frames, although they are correctly detected at all other frames. In this paper, we analyze the mechanism of how such miss-detection occurs. For the most popular class of detectors that are based on anchor boxes, we show the followings: i) besides apparent causes such as motion blur, occlusions, background clutters, etc., the majority of remaining miss-detection can be explained by an improper behavior of the detectors at boundaries of the anchor boxes; and ii) this can be rectified by improving the way of choosing positive samples from candidate anchor boxes when training the detectors.

## 1. Introduction

Detecting objects in an image is a fundamental problem of computer vision. The employment of convolutional neural networks (CNNs) has led to significant performance improvement on the task in recent years [17, 11, 16, 9, 25, 26, 6], and the accuracy seemingly comes close to the upper bound of the task. However, when applying them to a video, there often emerge cases where an object is momentarily miss-detected, i.e., undetected exceptionally at a few frames, although it is successfully detected for all other frames [24, 28, 22]; examples are shown in Figs. 1 and 2.

In this paper, we attempt to understand the mechanisms of how such miss-detection, which we will call momentarily missed detection (MMD), occurs. While in some cases it is apparent what causes MMD, e.g., motion blur and occlusion, as shown in Fig. 2, in others it is unclear, as in the example of Fig. 1. We call the former *external* factors and the latter *internal* factors.

To the authors' knowledge, there is no study in the lit-



Figure 1. An example of momentarily missed detection of an object. The person riding a bicycle is continuously detected in a sequence of frames but is not detected at the frame in the middle. Unlike those in Fig. 2, there is no apparent cause explaining the miss-detection.



Figure 2. Examples of the miss-detection cases for which we can see apparent causes, i.e., *external* factors. The white boxes indicate the ground-truth boxes and the red ones with a score indicate the predictions with the highest scores amongst those with  $\text{IOU} > 0.5$ . From left to right, the miss-detection is considered to occur due to motion blur, occlusion, and cluttered background.

erature that shows detailed analyses as to how many out of such miss-detection cases are attributable to external factors and internal factors, and what mechanisms are behind the internal factors. This may be reasonable, since MMD by definition occurs only rarely and thus has only small impact on detection accuracy that is usually measured by mAP (mean average precision).

In this paper, aiming to answer these questions, we conduct a series of analyses on the class of detectors that use anchor boxes, which is currently the most popular approach. In our experiments, we analyzed how SSD [11] and M2Det [26] work on 73 videos of diverse scenes contained in the DAVIS dataset [15]. We then found that MMD occurs at a few percent of about 8,000 frames in these videos, for which we will show the followings:

- About 60 to 70% of the MMD cases are attributable to external factors, such as motion blur, occlusion, background clutter, etc.
- About 20 to 30% of the MMD cases, for which there is no apparent cause like the external factors, can be explained by an improper behavior of the detectors at the boundaries of anchor boxes. Specifically, when an object moves in a video, its predicted score can drop considerably at the very instant when the optimal anchor for the object is switched from one to its neighbor.
- This behavior can be rectified by improving the method of choosing positive samples out of candidate anchors when training the detectors. This prevents the occurrence of MMD in most of the above cases without sacrificing overall detection accuracy; it also contributes to reduce MMD cases caused by the external factors.

As MMD occurs only rarely, its removal contributes only a little to improving detection accuracy that can be measured by mAP. Then, a practical value of our study, in addition to deeper understanding of how the detectors work, is as follows. The major application of object detection is arguably tracking moving objects in a video. To use a CNN-based detector for this purpose, we will first apply it to each video frame independently and then associate the detection results over the frames. Before the emergence of powerful CNN-based detectors, the second step of data association was relatively important, as the first step can only be performed with limited accuracy. The recent CNN-based detectors achieve much higher accuracy, which is sometimes nearly perfect. This lowers the relative importance of the second step, except for some hard cases, e.g., when many moving objects overlap with each other. If we can reduce the occurrence of MMD toward zero, we may be able to further accelerate this trend.

## 2. Related Work

In recent studies, object detection is posed as a regression problem, in which the geometry of an object bounding box in an image is predicted. There exist a few approaches of formulating this regression problem. The major approach is to use *anchor boxes*, which are default bounding boxes

with several predefined sizes and aspect ratios that are positioned at each grid cell of a feature map. The problem is then formulated as prediction of offsets to the true bounding box from its closest anchors, making it easier to deal with bounding boxes which usually have high degrees of freedom. The methods based on this formulation are further divided into two categories. One uses multiple feature maps in different resolutions [11, 3, 8, 9, 25, 26] and the other use a single feature map [17, 16]. These two strategies are compared in [8, 14].

For these methods, in this paper, we analyze momentary miss-detection that is often observed when applying them to a video sequence. To the authors' knowledge, there are few studies on this issue in the literature. An exception we are aware of is the study of Zhang *et al.* [24], in which considering the situation where a single-image object detector is applied to a continuous sequence of a video, they propose a new measure evaluating temporal (in)consistency of detection results and show that there is a trade-off between accuracy and stability. Our results in this paper do not disagree with theirs but show that we can improve the trade-off. It should also be noted that there are a number of studies that consider video object tracking, such as [28] and [22], in which the input is a set of multiple frames and the output is the trajectories of objects.

Recently, several methods have been proposed that do not use anchor boxes, such as CornerNet [6] and CenterNet [1]. Instead of predicting the offset from anchor boxes, they directly predict key points associated with bounding boxes, such as their top-left and bottom-right corners. Tian *et al.* [19] propose a method that directly predicts the location of an object at each grid point of a feature map, which is inspired by FCN-based architectures [5]. Although these approaches seem promising, it is too early to say that they will fully replace anchor-based methods. It is noteworthy that several researchers have proposed to use additional components to improve the anchor-based methods. Yang *et al.* [23] propose a method that dynamically generates appropriate anchor boxes from the input by using an additional network called the anchor function generator. Zhu *et al.* [27] propose to choose the best one of multi-scale feature maps for each input and predict bounding boxes at its grid points, which is integrated with predictions from the standard anchor-based method.

## 3. Momentarily Missed Detection

We are interested in cases where detection of an object fails at a particular frame of a sequence and is successful for all the other frames in the sequence. We call such miss-detection *momentarily missed detection* (MMD).

### 3.1. Definition

To conduct systematic analyses, we declare the frame  $t$  to be a MMD frame if it satisfies the following conditions:

$$p_{t-1}^c \geq \gamma_{min} \text{ and } p_{t+1}^c \geq \gamma_{min}, \quad (1a)$$

$$p_t^c / p_{t-1}^c \leq \gamma_{ratio}, \quad (1b)$$

$$p_t^c < \gamma_{max}, \quad (1c)$$

where  $p_t^c$  is the score of the object of class  $c$  that is given by our detector for frame  $t$ . Rigorously, we select the anchor box providing the largest score from those having  $\text{IOU} > 0.5$  with its ground truth box;  $p_t^c$  is the score for the anchor. The first condition (1a) ensures that the last and next frames (i.e.,  $t-1$  and  $t+1$ ) provide sufficiently high detection score. The second one (1b) requires that detection score drop at this frame  $t$  from the last frame  $t-1$ , at least to a certain extent. The last one (1c) eliminates frames at which the detection score is very high although the first two conditions are met.

When a detector detects objects successfully (except for MMD frames), the score  $p_t^c$  defined as above is usually greater than 0.7 (for any  $t$ ), and thus (1c) with  $\gamma_{max}$  having a lower value with a certain gap suffices to find most of the MMD frames of our interest. Thus, we set  $\gamma_{max} = 0.6$  in our experiments. However, there are additional cases where the score is continuously close to (but greater than) a natural threshold 0.5 and exhibits an exceptionally lower score at a single particular frame  $t$ . For completeness, we want to extract such frames as well, employing (1a) and (1b); we set  $\gamma_{min} = 0.5$  and  $\gamma_{ratio} = 0.9$  in our experiments.

### 3.2. Factors Causing MMD

#### 3.2.1 External Factors

We now consider why MMD occurs, i.e., why an object that is continuously detected suddenly ceases to be detected at a particular frame. The most likely cause will be bad imaging conditions that momentarily emerge at that frame, such as motion blur, illumination changes, occlusion by other objects, cluttered background, etc. Figure 2 shows typical examples, which are selected from those automatically selected by the conditions (1). It should be noted that several solutions to these cases are proposed in [10, 12, 20]. We refer to these causes as *external* factors.

#### 3.2.2 Anchor Boundary

However, there are cases where we cannot find such apparent causes, as shown in Fig. 1. What causes MMD in such cases? We conjecture it is attributable to (suboptimal) use of anchor boxes.

We are considering detectors that use anchor boxes. They are created at each grid point of a feature map with

a number of sizes and aspect ratios, aiming at easing regression of bounding boxes with high degrees of freedom. Thus, there are usually a large number of anchor boxes with various locations, sizes, and aspect ratios. However, they are only sparsely sampled in these parameter space. Thus, when a moving object continuously changes its location, size, and shape in a video, the optimal anchor that is responsible for its detection will be switched from one to its neighbor, as shown in Fig. 3. It is noteworthy that as the fewer the anchors are, the faster the overall computation is, it is not reasonable to use an excessive number of anchors.

Based on this structure of the detectors, we introduce the following hypothesis as to how MMD emerges that cannot be explained by the external factors:

**Hypothesis 1** *MMD can emerge at the boundaries of two neighboring anchors having either different scales, locations, or aspect ratios.*

We will experimentally validate this hypothesis. To do so, we first propose a method for analyzing the behavior of detectors at around anchor boundaries.

## 4. Analysis and a Solution

### 4.1. Understanding Behaviors of a Detector

#### 4.1.1 Image Warp Simulating Object Motion around Anchor Boundaries

Suppose we find MMD that occurs at an image of a sequence and wish to analyze how our detector behaves on that image and its neighbors in the sequence. To do this, we simulate the image motion of the object by warping the image at which the MMD occurs and examining detector behaviors on the warped images.

To be specific, we first apply a series of geometric warp to the image, generating image sequences, in each of which the object appears in either different sizes, positions, or aspect ratios. Then, the optimal anchor for the object will be switched from one to another in the images in each sequence, as shown in Fig. 4. Next, we run the detector on each image sequence and examine its outputs. *If MMD does occur at a boundary of anchors, we should be able to observe a decrease in the predicted score for the object at around the switch of the optimal anchor.*

To consider the three types of anchor boundaries shown in Fig. 3, we consider scaling, horizontal shift, and change in aspect ratio for the image warping. Their details are as follows:

**Scaling:** The image is enlarged and shrunk by a factor of  $1.02^n$  and  $0.98^n$  respectively, for  $n = 1, \dots, 29$ , yielding 59 images including the original.

**Shifting:** We consider only horizontal shift. The image is shifted in the  $x$  axis by  $3n$  pixels, for  $n = -29, \dots, 29$ ,

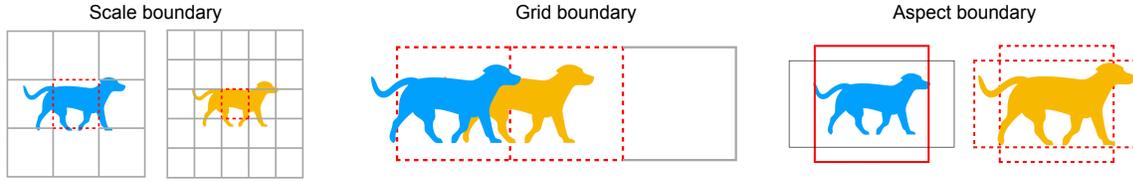


Figure 3. Three types of boundaries between neighboring anchor boxes.

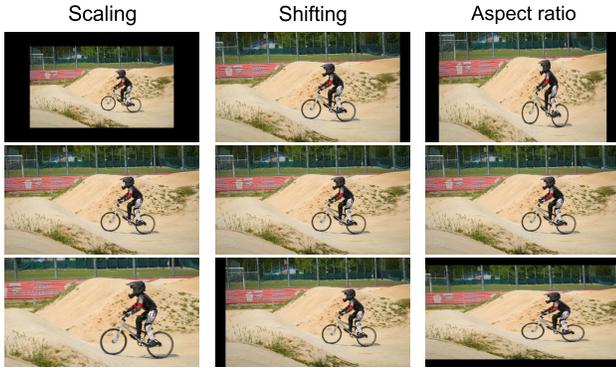


Figure 4. Examples of the transformed images by scaling, horizontal shift, and aspect ratio (from left to right). The top and bottom rows show the transformed images at  $n = \pm 15$ . If an object is miss-detected in the original image because the object size is on the boundary of anchors, then these transformed images will break the balance, making detection of the object back to normal.

yielding 59 images in total.

**Aspect ratio:** The image is enlarged and shrunk in either of  $x$  or  $y$  axis by a factor of  $1.01^n$  and  $0.99^n$ , respectively, for  $n = 1, \dots, 29$ , yielding 59 images in total for each axis.

#### 4.1.2 Case Studies

How does a detector behave around an anchor boundary? A typical behavior is shown in Fig. 5. The plot shows how the score for an object varies with respect to its size change, which is simulated by scaling the image. The blue crosses indicate the scores predicted for an anchor created on the  $19 \times 19$  feature map, whereas the magenta stars indicate those for an anchor on the  $10 \times 10$  feature map. It is seen from Fig. 5 that the former gives higher scores in the negative ( $n < 0$ ) region (i.e., smaller object size), whereas the latter gives higher scores in the positive ( $n > 0$ ) region (i.e., larger object size), indicating that the optimal anchor is switched at around the original object size ( $n = 0$ ). It is also seen that the detection score becomes lower than it should be at around the original size ( $n = 0$ ). Ideally, the score should be kept high throughout a range centered at  $n = 0$ , resulting in stable detection of the object. The lower

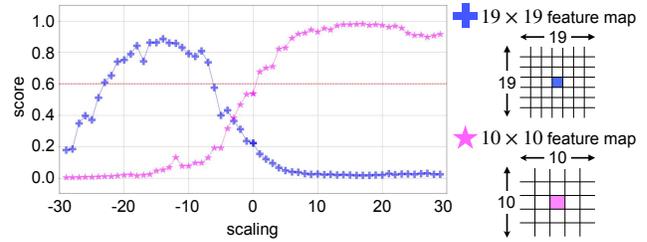


Figure 5. An example of how a detector (SSD) behaves around a boundary of two anchors with different sizes (scales). The crosses and stars indicate the score of two neighboring anchors, one from  $19 \times 19$  and the other from  $10 \times 10$  feature maps, respectively, for a sequence of images that are obtained by scaling the original image; the horizontal axis indicates the scaling factor, e.g.,  $1.02^n$ .

scores at around  $n = 0$  provide a direct explanation as to why MMD occurs at the original image.

As shown in Fig. 6, a similar behavior is observed for boundaries of anchors defined on different grid points. The blue crosses and the magenta crosses show detection scores for two neighboring anchors,  $(x, y) = (11, 9)$  and  $(12, 9)$ , respectively, which are created on the same  $19 \times 19$  feature map. The former gives higher scores in the negative region (i.e., shifting the image to the left) and the latter gives higher scores in the opposite region (shifting to the right). The optimal anchor is switched at around the center, where the scores for the two anchors are both lower than they should be. This explains why MMD occurs at the original image.

We will show the results of experiments conducted on a large number of images to validate Hypothesis 1 in Sec. 5.

#### 4.2. Improving Detector Behavior around Anchor Boundaries

Assuming Hypothesis 1 to be true, how can we improve the behavior of detectors around anchor boundaries? We conjecture that the drop of detection score on anchor boundaries, as seen in Figs. 5 and 6, is attributable to sub-optimal methods for selecting positive samples. We then present an improved method.

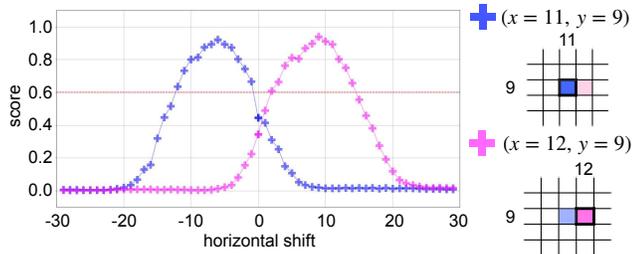


Figure 6. An example of how a detector (SSD) behaves around a boundary of two anchors on different grid points. The blue and magenta crosses indicate the score for two neighboring anchors, one from the grid point  $(x = 11, y = 9)$  and the other from  $(x = 12, y = 9)$  in the  $19 \times 19$  feature map, respectively.

Table 1. Various strategies to select positive (and negative) samples in popular detectors, which are either from the original papers or from authors’ implementations.  $C_{ij}$  in YOLOv2 is an indicator showing if the center of an object is on the grid  $i$  and its associated anchor  $j$  achieves the highest IOU with the ground-truth boundary box.

models	positive	negative	anchors	HNM
Faster R-CNN [17]	IOU > 0.7	IOU < 0.3	9	-
SSD [11]	IOU > 0.5	-	4 or 6	1:3
RetinaNet [9]	IOU ≥ 0.5	IOU < 0.4	9	-
RefineDet [25]	IOU > 0.5	-	4	1:3
M2Det [26]	IOU ≥ 0.5	-	6	1:3
YOLOv2 [16]	$C_{ij} = 1$	$C_{ij} = 0$ and IOU ≤ 0.6	5	-

#### 4.2.1 Conventional Method for Positive Sample Selection

To achieve high recall, many recent detectors has a common design that a CNN first generates a number of candidate boxes for a single object, and they are all discarded but a single box in a post filtering process, such as non-maximum suppression and center-ness scores [19]. Thus, when training the CNN, multiple anchor boxes are selected as positive samples; that is, we train the CNN to declare the presence of the object class for any of the selected anchor boxes. In many popular detectors [17, 11, 9, 25, 26], the selection of these anchors is performed by simple thresholding of IOU between the default anchor box and the ground truth box, as shown in Table 1.

An apparent issue with such binary thresholding is that only a slight difference in IOU leads to opposite results; for instance, an anchor with IOU= 0.501 is chosen as a positive sample, whereas that with IOU= 0.499 is not. Our conjecture is that this coarse selection of samples leads to the aforementioned improper behaviors of detectors.

#### 4.2.2 Proposed Sampling Method

We propose to make a softer decision for the selection of positive samples. The basic idea is to incorporate a con-

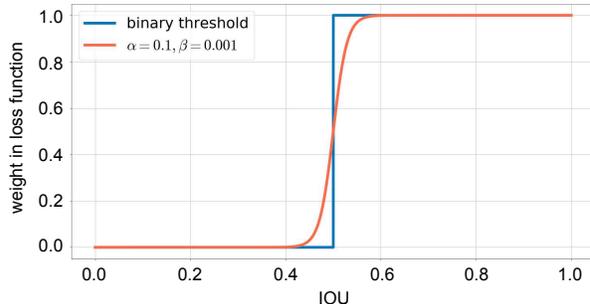


Figure 7. Incorporated weight  $X'_k$  applied on the loss of anchor  $k$  having IOU  $r$  with the ground truth box ( $\alpha = 0.1$  and  $\beta = 0.001$ ). Many previous methods employ binary thresholding with IOU = 0.5 to choose positive samples. Our weight performs soft thresholding.

tinuous weight tied with the IOU between the anchor box and the ground truth box into the evaluation of the loss. In a standard implementation of the conventional sampling scheme, an indicator variable  $X_k \in \{0, 1\}$  is used to express whether the  $k$ -th anchor box will be chosen as a positive sample or not, which is computed by thresholding the IOU, and then it is used as a weight of the loss for this anchor box. We extend  $X_k$  to the continuous domain, i.e.,  $X'_k \in [0, 1]$ ;  $X'_k$  is multiplied with the loss for the  $k$ -th anchor as in the conventional method.

We compute the new indicator  $X'_k$  using the IOU  $r$  between the anchor  $k$  and the ground truth bounding box as

$$X'_k = f(r), \quad (2)$$

where  $f$  is a logistic sigmoid function with a parameter  $a$ :

$$f(r) = \frac{1}{1 + \exp\{-a(r - 0.5)\}}. \quad (3)$$

For the sake of computational efficiency, we consider this soft thresholding only in the range  $[0.5 - \alpha, 0.5 + \alpha]$  of IOU  $r$ . That is, we simply set  $X'_k = 0$  for  $r < 0.5 - \alpha$  and  $X'_k = 1$  for  $r > 0.5 + \alpha$ . To smoothly connect  $X'_k$  at the borders  $r = 0.5 \pm \alpha$ , we set  $a$  so that  $f(0.5 - \alpha) = \beta$  (or equivalently,  $f(0.5 + \alpha) = 1 - \beta$ ) for a small  $\beta$ . We set  $\alpha = 0.1$  and  $\beta = 0.001$  throughout our experiments. To avoid the case where no anchor is assigned to a ground truth box, we follow previous studies, setting  $X'_k = 1$  for the anchor  $k$  with the largest IOU, when no anchor with  $X_k > 0$  exists.

## 5. Experimental Results

### 5.1. Detectors

For detectors, we choose SSD [11] and M2Det [26], which are popular two detectors that are based on anchor boxes. For the sake of experimental reproducibility,

we used their implementation that are publicly available, i.e., SSD-VGG16<sup>1</sup>, SSD-ResNet50<sup>2</sup>, and M2Det-VGG16<sup>3</sup> (VGG16/ResNet50 indicates a backbone network [18, 4]). We trained them on the most popular dataset for object detection, PASCAL VOC.

**Details of training** The PASCAL VOC dataset [2] 2007+2012 trainval split is used for training. SGD with momentum was used for the optimizer, where momentum was set to 0.9 with weight decay  $5.0 \times 10^{-4}$  in all models. In the training of SSD [11], setting the initial learning rate to  $1 \times 10^{-3}$ , we decrease it to  $1 \times 10^{-4}$  at 120 epochs and  $1 \times 10^{-5}$  at 160 epochs; training is stopped at 200 epochs. In the training of M2Det [26], initial learning rate is set to  $4 \times 10^{-3}$ , it is decreased to  $2 \times 10^{-3}$ ,  $4 \times 10^{-4}$ ,  $4 \times 10^{-5}$ , and  $4 \times 10^{-6}$  at 40, 60, 80, and 100 epochs; training is stopped at 120 epochs. For M2Det, weights are initialized by a model pretrained on COCO dataset [7] and finetune them.

## 5.2. Dataset

We choose DAVIS (Densely Annotated Video Segmentation) 2017 dataset [15] for our experiments. Existing datasets for object detection such as PASCAL VOC [2] and COCO [7] contain only still-images and cannot be used. Datasets for object tracking such as [21, 13] do contain videos but tend to lack a variety of objects and changes in object size, aspect ratio, etc. The DAVIS dataset is ideal for our purpose, since it consists of 90 videos of various scenes and pixel-wise segmentation masks are provided for various objects in each frame of all the video. We automatically generate ground truth bounding boxes by obtaining the circumscribing box to each segmentation mask of an object.

As they are trained on PASCAL VOC, our detectors are trained to detect the 20 classes of objects contained in the dataset. We identify 14<sup>4</sup> of them in the videos of the DAVIS dataset, and thus use these objects for our experiments. We can use only 73 videos out of 90 and discard others that do not contain the 20 object classes. As there are often multiple objects (that are annotated) in an image and our analysis is done for each object, the number of video frames we used in our experiments amounts to 8, 140.

**Extraction of MMD Frames** We run each of the above detectors on the above data and then applied the criteria (1) to the detection results to extract target frames for which MMD occurs. This yields several hundreds of frames for each detector; exact numbers are shown in the row of

<sup>1</sup><https://github.com/qfgaohao/pytorch-ssd>

<sup>2</sup><https://github.com/ShuangXieIrene/ssds.pytorch>

<sup>3</sup><https://github.com/qijiezhao/M2Det>

<sup>4</sup>aeroplane, bicycle, bird, boat, bus, car, cat, cow, dog, horse, motorbike, person, sheep, and train

Table 2. Number of MMD frames found on selected video frames (8,140 images in total from 73 videos) of the DAVIS dataset for different detectors, and their classification into three categories. <sup>†</sup>mAP for each detector is evaluated on PASCAL VOC 2007 test set.

Models	SSD-VGG16	SSD-ResNet50	M2Det-VGG16
mAP <sub>0.5</sub> <sup>†</sup>	76.5	74.8	79.3
MMD frames	367	333	264
External factors	262	239	160
Anchor boundary	<b>73</b>	<b>61</b>	<b>75</b>
Others	32	33	29

‘MMD frames’ of Table 2. It is seen that the these frames are less than 5% of the total number of frames (i.e., 8, 140).

## 5.3. Classifying MMD Frames by Causes

We then visually inspect each of these MMD frames. We first check if the MMD at the frame is caused by an external factor, such as motion blur, occlusion, and cluttered background, as discussed in Sec. 3.2.1. There are frames at which the object is correctly detected but its predicted class is wrong, e.g., a dog is detected but is recognized as a cat; there are about 5 to 10 frames in total for each detector. We classified these frames also into the category of external factors. The row ‘External factors’ of Table 2 shows the number of these frames for each detector.

We next inspect each of the remaining MMD frames to judge whether it emerges due to the *anchor boundary* as is predicted in Hypothesis 1 or due to *other causes*. To do this, we apply the method in Sec. 4 to these MMD frames; this yields sequences of warped images, for each of which we run the same detector. We then plot detection scores outputted for the associated neighboring anchors, by using which we make the above decision. The rows ‘anchor boundary’ and ‘others’ of Table 2 show the number of MMD frames classified into the two categories, respectively.

These results are summarized in the bar plots of Fig. 8. It is seen that while the external factors are dominant, the hypothesized cause of anchor boundaries occupies non-negligible portion. It will be the most effective if we can cope with the external factors; for instance, it may be effective to remove motion blur from the input image in a pre-processing step before the application of an object detector. This is, however, not a simple task. On the other hand, as will be shown in the next subsection, we can reduce the number of MMDs caused by anchor boundaries *for free*, i.e., by simply switching the method for positive sample selection from the conventional one to the proposed method.

Note that there remain a portion of MMD frames that are not explained by the external factors or the anchor bound-

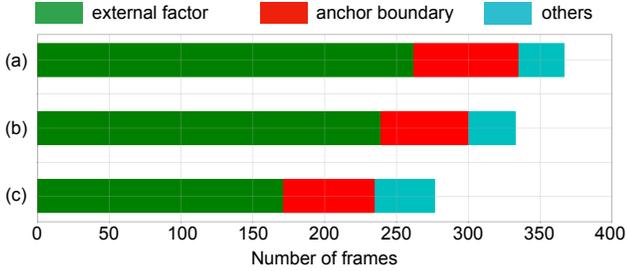


Figure 8. Bar plots of the same data shown in Table 2. (a) SSD-VGG16, (b) SSD-ResNet50, and (c) M2Det-VGG16.

aries. It should also be noted that we could not find a MMD frame caused by anchor boundaries in terms of aspect ratio. This may be because objects tend to change their aspect ratio less frequently than they change their sizes or exhibit translational motion.

#### 5.4. Evaluation of Proposed Sampling Method

We also conducted experiments to examine the effectiveness of the proposed method for selecting positive samples and weighting them. In the experiments we trained the same three detector models, SSD-VGG16, SSD-ResNet50, and M2Det-VGG16 using the proposed sampling/weighting method explained in Sec. 4.2. Their training in other parts is performed in the same way as before. We applied the trained models on the same dataset (i.e., 8,140 images from 73 DAVIS videos) and then analyzed the detection results using the same method as the previous experiments. The results are shown in Table 3. The bar plot of the same data including Table 2 are shown in Fig. 9.

It is seen from the table and the figure that the employment of the proposed sampling/weighting method decreases the number of MMD frames caused by anchor boundaries significantly; namely, from 73 to 11 with SSD-VGG16, from 61 to 14 with SSD-ResNet50, and from 75 to 19 with M2Det-VGG16.

Figure 10 shows a number of examples of how MMD is resolved by the incorporation of the proposed method. The left panel shows detection results of SSD-VGG16 with the conventional sampling method (i.e., binary thresholding), which suffer from MMD. Each row shows a pair of a MMD frame and the score profile for neighboring anchors over associated image warping. As is explained in Sec. 4.1.2, it is seen for each case that there emerges a valley in the score profile at around the anchor boundary, which well explains why MMD occurs at the frame. The right panel shows results on the same frames of the same detector trained with the proposed method. Each row shows a pair of the detection result and the corresponding score profiles for the neighboring anchors. In each case, the object that is missed on the left panel is correctly detected. It is seen for

Table 3. Comparisons between binary and soft thresholding for positive sample selection. We set hyperparameters for soft threshold as  $[\alpha, \beta]=[0.1, 0.001]$ .  $\dagger$ mAP values are evaluated on the PASCAL VOC 2007 test set.

Models	SSD-VGG16		SSD-ResNet50		M2Det-VGG16	
	Binary	Soft	Binary	Soft	Binary	Soft
mAP $_{0.5}^{\dagger}$	76.5	77.2	74.8	74.7	79.3	80.4
MMD frames	367	206	333	247	264	182
External factors	262	180	239	194	160	140
Anchor boundary	73	<b>11</b>	61	<b>14</b>	75	<b>19</b>
Others	32	15	33	39	29	23

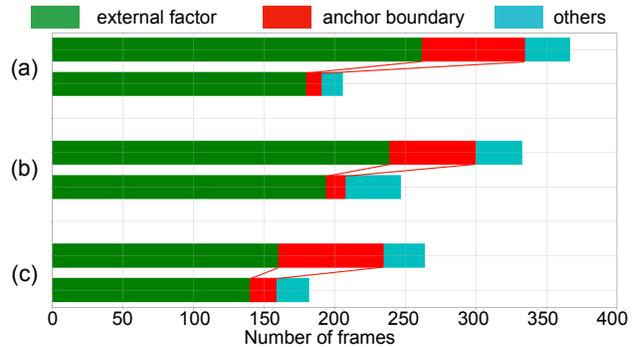


Figure 9. Bar plots of Table 3 (and Table 2 for the sake of comparison). (a) SSD-VGG16, (b) SSD-ResNet50, and (c) M2Det-VGG16. It is seen that the number of MMD frames, particularly those due to anchor boundaries, decreases.

each case that the peak of the score curve for each anchor tends to be more flat, resulting in shallower valley at the anchor boundary. This well explains the vanishing of MMDs in these examples.

We may conclude from these results the effectiveness of our approach. Moreover, this will also be an additional support for the validity of Hypothesis 1. Another remark with the results is that MMD frames caused by external factors also decreases to a certain degree; the largest decrease (more than 30%) is observed for SSD-VGG16. Although the mechanism behind this improvement is not clear, we can confirm the proposed method does not worsen the performance or even contribute to improvements overall. This is also confirmed by that mAP maintains the same level or even shows some improvements. As is mentioned earlier, we set  $\alpha = 0.1$  and  $\beta = 0.001$  in the experiments. It is noteworthy that we chose these values intuitively and there may be better values leading to even better performance.

## 6. Summary and Conclusion

Recent CNN-based detectors are designed to work on a single image and are usually trained using a large number of labeled still-images that are independent of each other.

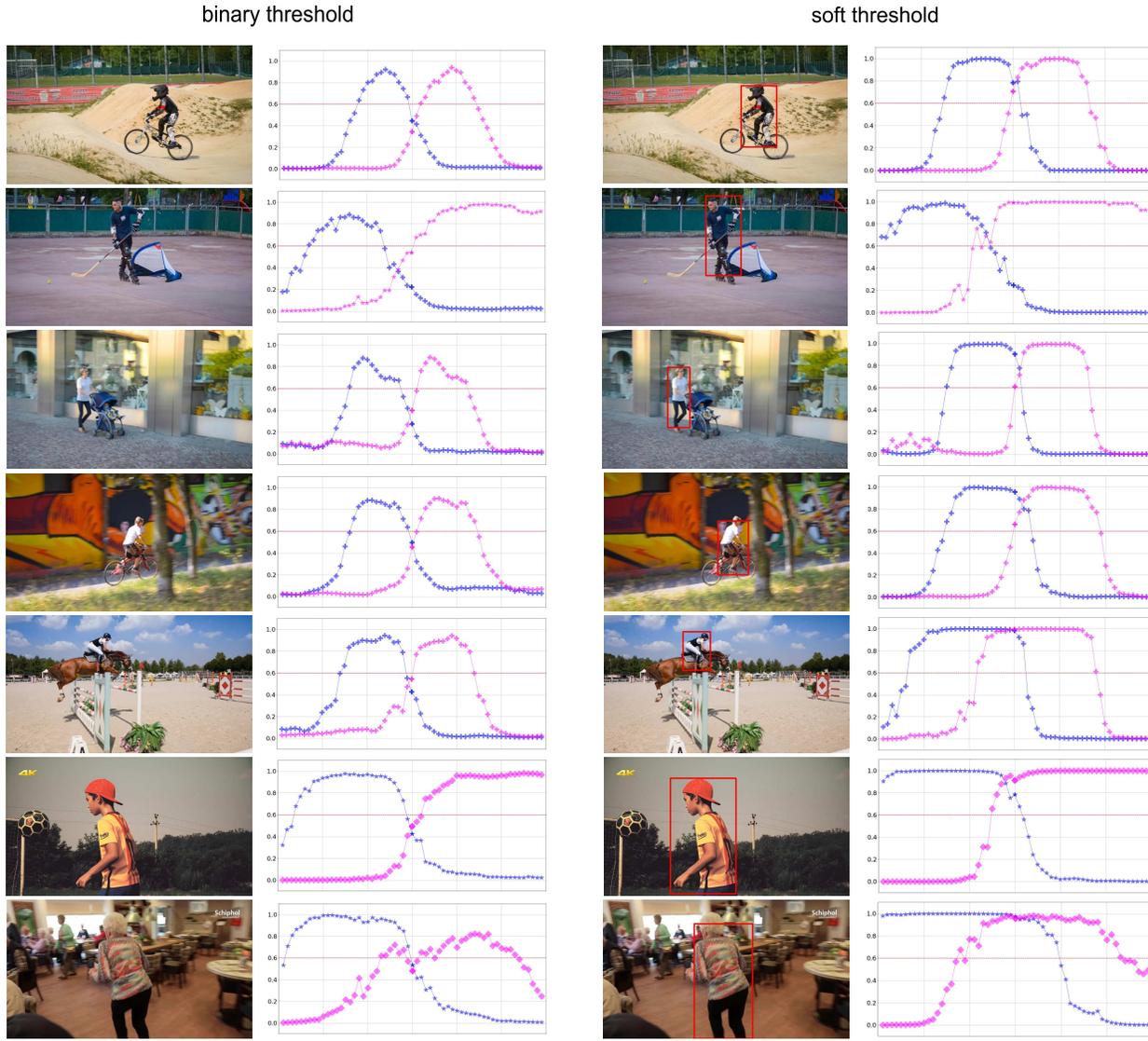


Figure 10. Examples of frames at which SSD-VGG16 trained with the conventional method for positive sample selection (i.e., binary thresholding) causes MMD, whereas the same model trained with the proposed method (i.e., soft thresholding) ceases to cause MMD. The plots on the second columns show the scores for two neighboring anchors, from which it is confirmed that the proposed sampling method contributes to fill in the gap at around anchor boundaries.

In this paper, we have analyzed momentarily missed detection (MMD) that is often observed when we apply these detectors to a sequence of video frames. We have revealed through several experiments that i) external factors (e.g., motion blur etc.) explains the majority of MMD cases; ii) the remaining MMD cases can be mostly explained by an improper behavior of the detectors at boundaries of anchor boxes; and iii) it can be rectified by using the improved method that chooses positive samples from candidate anchor boxes when training the detectors.

## Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number JP15H05919 and JP19H01110 and JST CREST Grant Number JPMJCR14D1.

## References

- [1] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. CenterNet: Keypoint triplets for object detection. In *Proc. CVPR*, 2019.
- [2] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC)

- Challenge. *IJCV*, 2010.
- [3] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. DSSD : Deconvolutional single shot detector. *arXiv:1701.06659*, 2017.
- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 2016.
- [5] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv:1509.04874*, 2015.
- [6] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proc. ECCV*, 2018.
- [7] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014.
- [8] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In *Proc. CVPR*, 2017.
- [9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *Proc. ICCV*, 2017.
- [10] C. Liu, P. Liu, W. Zhao, and X. Tang. Robust tracking and redetection: Collaboratively modeling the target and its context. *TMM*, 2018.
- [11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *Proc. ECCV*, 2016.
- [12] B. Ma, L. Huang, J. Shen, L. Shao, M. Yang, and F. Porikli. Visual tracking under motion blur. *TIP*, 2016.
- [13] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016.
- [14] X. Ming, F. Wei, T. Zhang, D. Chen, and F. Wen. Group sampling for scale invariant face detection. In *Proc. CVPR*, 2019.
- [15] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. Sorkine-Hornung, and L. V. Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*, 2017.
- [16] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *Proc. CVPR*, 2017.
- [17] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proc. NIPS*, 2015.
- [18] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. ICLR*, 2015.
- [19] Z. Tian, C. Shen, H. Chen, and T. He. FCOS: fully convolutional one-stage object detection. In *Proc. ICCV*, 2019.
- [20] R. Walsh and H. Medeiros. Detecting tracking failures from correlation response maps. In *Proc. ISVC*, 2016.
- [21] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *TPAMI*, 2015.
- [22] F. Xiao and Y. J. Lee. Video Object Detection with an Aligned Spatial-Temporal Memory. In *Proc. ECCV*, 2018.
- [23] T. Yang, X. Zhang, Z. Li, W. Zhang, and J. Sun. Metaanchor: Learning to detect objects with customized anchors. In *Proc. NIPS*, 2018.
- [24] H. Zhang and N. Wang. On The Stability of Video Detection and Tracking. *arXiv:1611.06467*, 2016.
- [25] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-Shot Refinement Neural Network for Object Detection. In *Proc. CVPR*, 2018.
- [26] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proc. AAAI*, 2019.
- [27] C. Zhu, Y. He, and M. Savvides. Feature selective anchor-free module for single-shot object detection. In *Proc. CVPR*, 2019.
- [28] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-Guided Feature Aggregation for Video Object Detection. In *Proc. ICCV*, 2017.