



Fast detection of bag-breakups in pulsating and steady airflow using video analysis and deep learning

Daiki Morita¹ · Bisser Raytchev¹ · Abdussalam Elhanashi² · Mikimasa Kawaguchi¹ · Yoichi Ogata¹ · Toru Higaki¹ · Kazufumi Kaneda¹ · Akira Nakashima³ · Sergio Saponara²

Received: 21 January 2023 / Accepted: 12 September 2023 / Published online: 16 October 2023
© The Author(s) 2023

Abstract

Object detection methods based on deep learning have made great progress in recent years and have been used successfully in many different applications. However, since they have been evaluated predominantly on datasets of natural images, it is still unclear how accurate and effective they can be if used in special domain applications, for example in scientific, industrial, etc. images, where the properties of the images are very different from those taken in natural scenes. In this study, we illustrate the challenges one needs to face in such a setting on a concrete practical application, involving the detection of a particular fluid phenomenon—bag-breakup—in images of droplet scattering, which differ significantly from natural images. Using two technologically mature and state-of-the-art object detection methods, RetinaNet and YOLOv7, we discuss what strategies need to be considered in this problem setting, and perform both quantitative and qualitative evaluations to study their effects. Additionally, we also propose a new method to further improve accuracy of detection by utilizing information from several consecutive frames. We hope that the practical insights gained in this study can be of use to other researchers and practitioners when targeting applications where the images differ greatly from natural images.

Keywords Object detection · Scientific and industrial applications · Real-time processing · Small-size datasets · YOLOv7 · RetinaNet

1 Introduction

Object detection is one of the major tasks in computer vision and recently, various deep learning-based models with improved accuracy and detection speed have been proposed (see [3, 6, 7, 12] for recent comprehensive surveys on object detection). However, in practice, there are various problems in applying deep learning-based object detection models, especially in special domains, where images represent specific scientific phenomena or are taken in industrial settings, etc., and it is generally difficult to secure a large amount of annotated data for such applications. The reason

for this is that in such cases (in contrast with natural scene images), usually, only experts can provide the annotations, which makes the annotation process very costly. Therefore, in such cases, securing large annotated datasets for training the data-hungry deep learning models is either impossible or impractical. Additionally, often the presence of rare classes of objects, or phenomena to be detected, leads to the class imbalance problem, which can make the learning process difficult and unreliable.

Furthermore, it is a common practice that newly developed and state-of-the-art object detection models are mainly evaluated using natural image-based datasets such as the PASCAL VOC (PASCAL Visual Object Classes Challenge) dataset [5] and the COCO (Microsoft Common Objects in Context) dataset [11]. Since the images in these datasets (and the features extracted from them) would differ significantly from those targeted in general scientific/industrial applications, it is unclear how accurate and effective the methods are, and which of their components are critical for the success of the targeted application. Sometimes, the practitioners who need to provide the solutions (using appropriate object

✉ Bisser Raytchev
bisser@hiroshima-u.ac.jp

¹ Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima, Japan

² Dip. Ingegneria Informazione, University of Pisa, Pisa, Italy

³ MBD Innovation Department, Mazda Motor Corporation, Fuchu, Japan

detection methods) are scientists or other staff which are not machine learning experts, and for such people, it can be very difficult to orientate themselves in the extremely rapidly evolving landscape of object detection algorithms, very often leaving them without a clue which method should they use and which strategies should they follow to obtain the best possible performance on their application.

Therefore, in this study, we apply two representative object detection models to a small-size dataset from a domain that differs significantly from natural images, and examine their detection accuracy and the effects of the models' components on accuracy. The dataset used in this study consists of basic experimental images representing a phenomenon called *droplet dispersion*, which occurs inside an automobile exhaust pipe. The dataset consists of a small number of about 800 images, and the target of detection is a form of droplet dispersal called *bag-breakup*. While in the case of natural scene images usually there can be dozens of detection targets in a single image, or it is easy to obtain a large number of images with the targets of interest (like cats, or cars, etc.), our dataset has typically one and maximum two detection targets in a single image, while we can have many thousands of images collected from the experiments, which do not contain any detection target at all. Additionally, what makes the droplet dispersion images more challenging is that the detection targets (bag-breakups in our task) are very similar to the background (in terms of visual appearance and texture patterns); while in natural scene images, it is much more easier to discriminate between, for example, cats and their background. Finally, in test mode, the trained model has to process an enormous volume of experimental data/images generated daily and input as a video stream to the system, therefore *real-time processing* abilities are crucial.

To address these issues, and provide guidance to practitioners who might have to deal with similar challenges, we consider a set of solutions and architecture choices suitable for such a setting and investigate their effectiveness. The rest of this paper is organized as follows. First, in the next section, we describe the concrete task which motivates our research. In Sect. 3, we discuss the most common problems and challenges which practitioners need to deal with when applying object detection methods in special domains and show how these can be overcome using corresponding practices and architecture decisions developed in recent state-of-the-art methods in the object detection field. Additionally, in Sect. 3.4 we propose a new method, which by utilizing information from several consecutive frames further improves accuracy of detection by eliminating false positives which might occur in regions which closely resemble the visual structure of bag-breakup patterns and are difficult even for experts to discriminate from a single image. Section 4 reports the experimental results, providing both quantitative

and qualitative assessment and finally Sect. 5 concludes the paper.

2 Bag-breakup

2.1 Motivation for the automatic detection of the bag-breakup phenomenon

In recent years, global warming and air pollution have become international problems [1, 21]. Automobile emissions contain not only greenhouse gases such as CO_2 , but also NO_x (nitrogen oxides), HC (hydrocarbons), CO (carbon monoxide), PM (particulate matter) and other harmful substances which pollute the air. Due to this fact, regulations on automobile emissions are becoming more stringent every year.

To reduce the amount of hazardous substances contained in emissions, it is important to burn fuel at a theoretical air–fuel ratio of (14.7 : 1), which is the ratio of air to fuel that allows the fuel to burn without excess or deficiency. To achieve this theoretical air–fuel ratio, O_2 sensors are installed in the exhaust pipes of automobiles. The O_2 sensor detects the oxygen concentration in the exhaust gas and precisely controls the air–fuel ratio to be close to the theoretical air–fuel ratio.

The temperature of the exhaust pipe wall is also lower in winter due to the lower ambient air temperature. When hot exhaust gas containing water vapor flows into the exhaust pipe, the water vapor is cooled and condensates on the exhaust pipe wall, causing water to accumulate inside the pipe. This condensate is released into the atmosphere as it is dispersed by the exhaust gas, and its adhesion to the O_2 sensor described above leads to deterioration of its control response performance, or in the worst case, damage to the sensor. In addition, in contrast to the case of condensate dispersing to the O_2 sensor, if the condensate does not disperse but accumulates in the exhaust pipe, due to freezing under sub-zero conditions, the exhaust pipe will not be able to carry away exhaust gases properly. For these reasons, it is important to elucidate the mechanism of condensate dispersion.

2.2 Experimental setting for the image acquisition

In our current research, we are using the experimental apparatus shown in Fig. 1 to take experimental images under various experimental conditions (wind speed, pipe shape, etc.) and visually confirm the frequency of droplet dispersal. However, droplet scattering occurs in only about $1/1000 \sim 1/10000$ of the experimental images, and having to visually inspect a huge number of images to detect droplet scattering is very time-consuming and burdensome.

Fig. 1 The experimental equipment used for image acquisition from the droplet dispersal experiments [14]

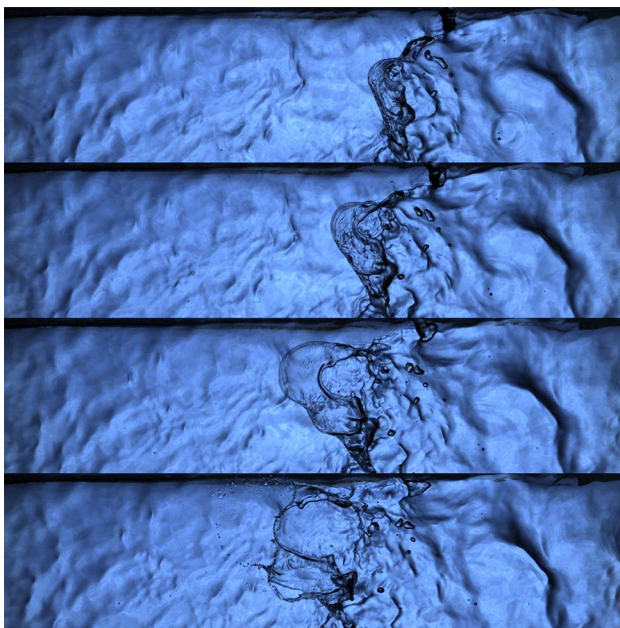
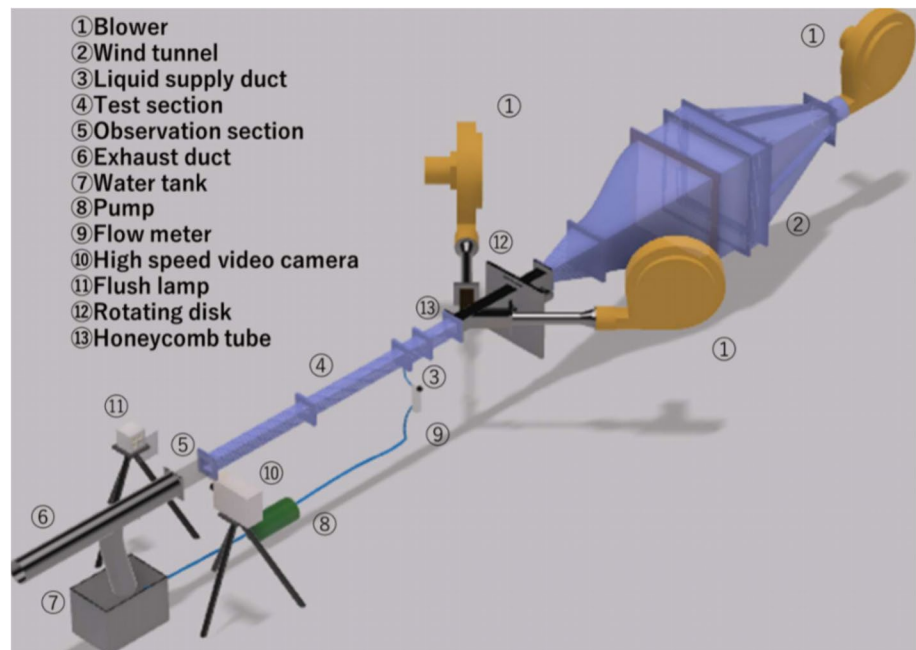


Fig. 2 An example of a bag-breakup (the bag can be seen moving from right to left)

Therefore, it is important to develop and implement a system that can automatically detect droplet scattering on behalf of the researcher. In this study, we attempt to solve this problem using an object detection method based on deep learning.

There are four types of scattering phenomena: ligament, bag-breakup, multimode, and impact. In this study, we concentrate on bag-breakup detection. An example of a bag-breakup is shown in Fig. 2. A bag-breakup is formed

when the liquid film is subjected to inertial forces from the gas phase, and the droplet expands like a balloon and then breaks up. It is called a bag-breakup because at the moment when the balloon breaks, it looks like a bag.

3 Problems and solutions for object detection in special domains

Here, we will describe the most common problems and challenges which practitioners would likely have to deal with when applying object detection methods in special domains (where the images and task constraints differ significantly from the natural images case), and will also show how these can be overcome using corresponding practices and architecture decisions (summarized in Fig. 3) developed in recent state-of-the-art methods in the object detection field.

3.1 Detection speed

In many industrial or scientific application settings, it is necessary to process an enormous volume of data given only a limited time span for processing, so *detection speed* is a crucial requirement for success in such cases. Generally, object detection methods can be divided into two groups: (a) *two-stage object detection models*, such as Faster-RCNN [16], in which a first region proposal stage (during which candidate object bounding boxes are generated) is separated from the second stage (during which a classification and regression tasks are performed in order to obtain object's class and corresponding bounding box); and (b) *one-stage*

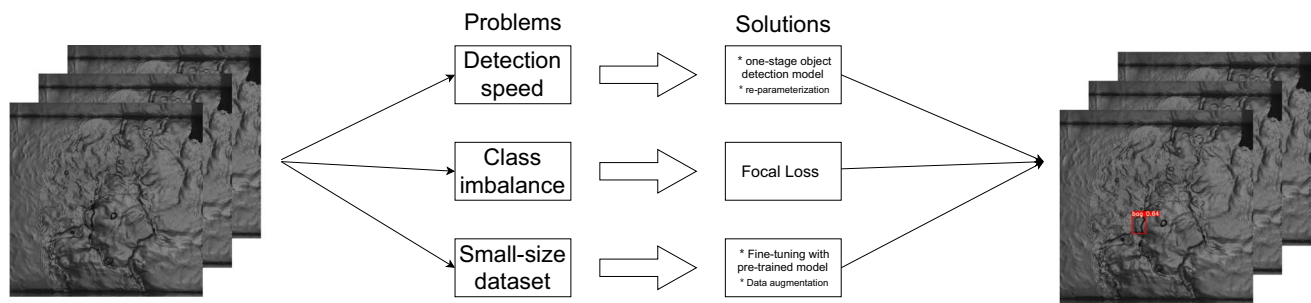


Fig. 3 Illustration of some of the most common problems and challenges which practitioners need to overcome when dealing with object detection in special domains and the general solutions attempted in the context of the present task (see main text for the details)

object detection models, such as SSD [13], YOLO [15, 17], EfficientDet [18] or recent transformer-based methods like [23], in which the objects' class and bounding boxes are predicted directly from the input images without the necessity of a preliminary region proposal stage. Because of the additional time required to compute the region proposals in two-stage models, one-stage models are generally much faster and therefore are a natural choice for applications like the current one, where detection speed is crucial.

Another area of research that has been active in recent years is module-level re-parameterization [4], where the multi-branch structure used during training is substituted by equivalent simpler modules during inference, thereby reducing memory access costs and speeding up the inference process. This approach is used in the state-of-the-art YOLOv7 model [19], and we will compare it in this study with the popular and widely used RetinaNet model [10], used as a baseline model here. RetinaNet is a simple but conceptually mature one-stage model, consisting of a backbone feature pyramid network (FPN) [9] and sub-networks for object classification and box regression. YOLOv7 is a one-stage model that utilizes various recent architectural design and learning techniques (using ELAN [20] as a backbone), which has shown superior performance to previous object detection models, both in terms of detection speed and accuracy.

3.2 Class imbalance

A challenge specific to the problem we deal in this paper (and industrial/scientific applications with similar settings) is that the object which needs to be detected represents a rare class/phenomenon, which can be observed only in a very small subset of all images, i.e., the predominant part of the images do not contain any target at all (i.e., represent the *background* class) and images on which the target can be seen contain usually only a single instance (a single object). This is very different from the typical case where the target objects (like cars, faces, etc.) appear in natural scenes. In

the latter case, usually a single image contains a multitude of objects and generally there are plenty of training images available (as it is easy to obtain and label them).

Since in our case most of the regions in a single image represent the background class, they would dominate the training, leading to the *class imbalance problem*, which can make the learning process difficult and unreliable. This problem is mitigated in the case of a two-stage detector because most of the background regions are eliminated during the region proposal stage, but it must be addressed in the case of a one-stage detector. A typical solution to this problem is the Focal Loss proposed by Lin et al. [10]. Focal Loss is defined by the following equation:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t).$$

By multiplying the usual Cross-Entropy Loss by $(1 - p_t)^\gamma$, the loss for samples with high confidence (corresponding to the predominant background class) becomes small, and training of the difficult samples (belonging to the target class which has much less training samples) can be emphasized. In this way, using the Focal Loss, RetinaNet although being a one-stage detector is able to achieve the same accuracy as a two-stage detector (but at the same time retaining the advantage of being much faster).

3.3 Small-size data sets

Another problem which needs to be addressed is that training of deep learning-based object detection methods typically requires huge training data sets, while for our task (in contrast with the general case where natural scene images are used) it is difficult to secure a large amount of annotated data, since only experts can provide the annotations, making this process very costly and thus impractical. To achieve high object detection accuracy while still training the model with only a small amount of data, fine-tuning is generally performed on the target data set, using weights that have been pre-trained on some large data set (usually

using readily available annotated images representing natural scenes) as initial values.

Data augmentation is also very effective for training with small data sets. In addition to the usual augmentation methods, such as image rotation, flipping, etc., recently special augmentation methods such as mosaic augmentation [2] have been used to combine multiple training images.

3.4 Utilizing information from neighboring frames

Although the strategies described above can be used successfully in most of the *general* problem settings targeted in this paper (i.e., in special domain applications where the images are very different from those taken in natural scenes), many *specific* problems might have additional information, which can be utilized to further improve the accuracy of detection. Here, we propose a novel method, which although being conceptually simple, using additional information specific to our target application (bag-breakup detection) is able to achieve higher accuracy than a state-of-the-art method like YOLOv7 used in the conventional way.

Our method is based on the observation that bag-breakup occurs across several consecutive frames and information from neighboring frames might be useful to detect difficult instances of bag-breakups which might be hard to detect from just a single frame (sometimes even human experts might not be able to accurately detect such difficult cases from a single image).

The main idea is illustrated in Fig. 4. Instead of training a model to detect a bag-breakup from a single image (say T -th frame in a sequence of images, as shown in the figure), as is the conventional way, we consider the $2i + 1$ frames in the interval $(T - i, \dots, T + i)$. We randomly select $(N - 1)/2$ frames among the frames in the interval $(T - i, \dots, T - 1)$

and also randomly select $(N - 1)/2$ frames among the frames in the interval $(T + 1, \dots, T + i)$ and stack them together with the T -th frame across the channel dimension. Note that N is an odd number and Fig. 4 illustrates the case for $N = 3$, i.e., a single frame somewhere *before* the T -th frame and a single frame somewhere *after* the T -th frame are stacked together with it. In order not to increase the channel dimension too much (and since color is not very informative for our task) we first convert the color images into grayscale. In this way, we obtain an input sequence of N images (or a single data sample consisting of N channels) which is input in the model (either RetinaNet or YOLOv7 in our case) and the model is trained or fine-tuned to detect all the bag objects in the image corresponding to the T -th frame.

The random selection of frames in the interval $(T - i, \dots, T + i)$ is used while training the model to make the method more robust to the different speeds of motion of the bags across different experiments (since it is necessary to detect bag-breakups in fluid flows moving at different speeds, e.g., when the wind speed is changed across different experiments).

In the next section, we show that our method significantly improves the detection accuracy compared with the conventional use of pre-trained state-of-the-art models and even enables a modified version of RetinaNet using our method to outperform YOLOv7 on our task.

4 Experimental results

4.1 Dataset

Previous studies have reported that droplets have different morphologies depending on conditions such as the

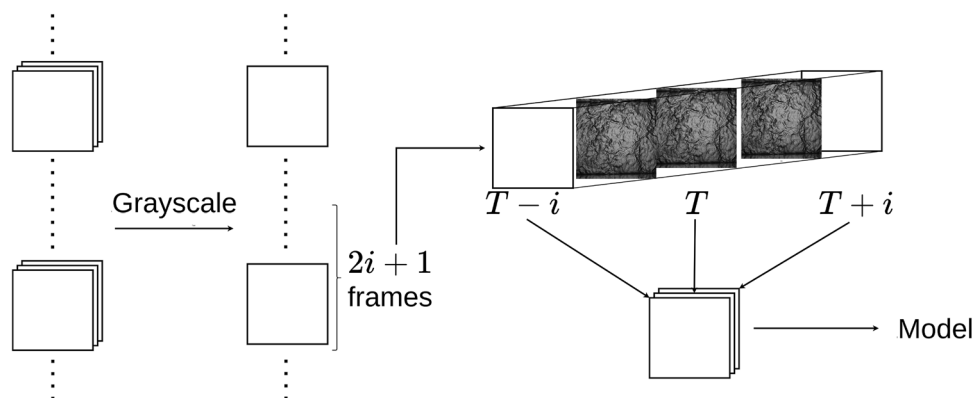


Fig. 4 Overview of the proposed method to use information from neighboring frames. To detect an object of interest in the T -th frame, we randomly select $N - 1$ frames among the frames in the interval $(T - i, \dots, T + i)$ and stack them together with the T -th frame across the channel dimension. The figure illustrates the case for $(N = 3)$, i.e.,

a single frame *before* the T -th frame and a single frame *after* the T -th frame are stacked together with it and features are extracted simultaneously from all the stacked frames. The random selection is needed to make the method more robust to the different speeds of motion of the bags across different experiments

magnitude and variation of wind velocity, and that they scatter randomly in time and space. Thus, they need to be statistically processed [14]. For the experiments reported below we used images from 72 different movies taken under 21 different experimental conditions (for example, different wind speed, etc.). All images were of size 1024×1024 pixels, and only 824 annotated frames were used for training and validation (since the annotation process is very time-consuming and costly).

4.2 Evaluation metrics

Average precision (AP) was used as an evaluation metric for the experimental results. AP is obtained by interpolating the precision–recall (PR) curve of the detection results and calculating the area of the region below it.

First, each detected object area which has a higher Intersection over Union (IoU) score (overlap ratio between detected and ground truth bounding boxes) than a pre-set threshold ϵ is matched with the corresponding ground truth bounding box. Next, the detected region with the highest confidence (among all detected regions associated with a certain ground truth bounding box) is then counted as a true positive (TP) if its confidence level is higher than some confidence threshold β , while all other detected regions are counted as false positives (FP). Therefore, each detection is either TP or FP, and by counting their number, a precision–recall curve can be obtained while varying the confidence threshold β . Finally, AP is calculated as the area under the precision–recall curve obtained in this way.

4.3 Model training

For our task, we trained two representative one-stage object detection models: the recently proposed YOLOv7 [19], representative of the current state-of-the-art in object detection, and the widely used RetinaNet [10] as a baseline. The data from all available source videos was divided into train, validation and test sets for evaluation. The number of images for train, validation and test were 597, 65 and 146 respectively. The images were resized to 800×800 pixels during training and validation.

In RetinaNet, Adam [8] was used as the optimization method. The learning rate was set to 1×10^{-5} , while β_1 and β_2 were set to $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size was set to 8. For data augmentation, vertical flip and color jitter were used.

In YOLOv7, only horizontal flip was changed to vertical flip for the data augmentation, while all other default settings were kept unchanged. Both models were trained for 100 epochs.

4.4 Experimental results

4.4.1 Model comparison

To prepare training data for object detection tasks, it is necessary that human annotators describe the exact location (i.e., provide the coordinates of the bounding box) of the objects of interest in the training images. Because bag-breakups can have very complex shapes (in terms of both texture patterns and visual appearance), unlike objects in natural scene images, it is difficult to determine the exact bounding box and there are large differences in annotations among individual annotators. For example, in natural scene images, all human annotators would agree what are the correct boundaries of a car or an animal, etc., while in the case of images like those we have to deal with in the present task, the boundaries of the bag-breakup objects are not precisely defined and different annotators would typically annotate the same object not exactly in the same way. Figure 5 shows an example where although the prediction of the model is very good, the Intersection over Union (IoU) score (i.e., the overlap ratio between detected and ground truth bounding boxes) between the bounding box predicted by the model and the annotated ground truth is quite low. For this reason, even though a typical value of the IoU threshold ϵ for natural scene images is 0.5, we also consider a lower one of 0.2, which might be sufficient for our task. The reason for this is that for the detection of objects in natural scene images usually a high overlap between a detection and ground truth of a car is natural to expect, in our case due to the large

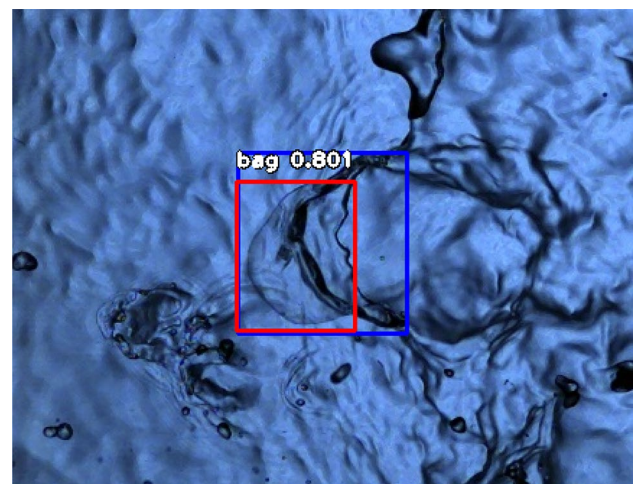


Fig. 5 Bag-breakup detection in an image, showing an example where even though the detection is very good, IoU is low. Model's prediction is shown in red and the ground truth annotation in blue. The number above the detected bounding box is the model's confidence

differences in annotations among individual annotators it is not reasonable (or necessary) to expect such a high overlap.

The evaluation results showing the average precision (AP) for each model are shown in Table 1. For this experiment, the RetinaNet model uses ResNet-34-FPN as a backbone, while the YOLOv7 model uses the regular YOLOv7 implementation [19]. The results in Table 1 show that changing the IoU threshold during evaluation improved APs by 10.1% for RetinaNet and 4.9% for YOLOv7. Also, as expected, in this experiment the state-of-the-art YOLOv7 model significantly outperforms RetinaNet by 8.9% for $\epsilon = 0.5$, which is

Table 1 Average precision (AP) obtained for both models for different values of the IoU threshold ϵ

Model	AP
RetinaNet (IoU=0.5)	0.817
RetinaNet (IoU=0.2)	0.918
YOLOv7 (IoU=0.5)	0.906
YOLOv7 (IoU=0.2)	0.955

reduced to 3.7% for the lower IoU threshold of $\epsilon = 0.2$. This indicates that the IoU threshold may need to be lowered during the evaluation of data such as those used in this study, where it is difficult to determine the exact bounding box of the target object and the differences among annotators are large.

More detailed evaluation results including the precision–recall curve and showing how precision, recall, and the F1 metric (which summarizes simultaneously both precision and recall) change as a function of each model's confidence are shown in Fig. 6. Figure 6 shows that YOLOv7 achieves higher accuracy in both precision and recall than RetinaNet and also has smaller changes in accuracy when the IoU threshold is changed. This may be due to the fact that YOLOv7 includes the loss related to IoU in the loss for the bounding box regression, while RetinaNet only uses smooth L1 Loss in that loss. However, from this figure, it can be seen that if the IoU threshold is lowered, as we suggest for this task (and similar tasks) the difference between using

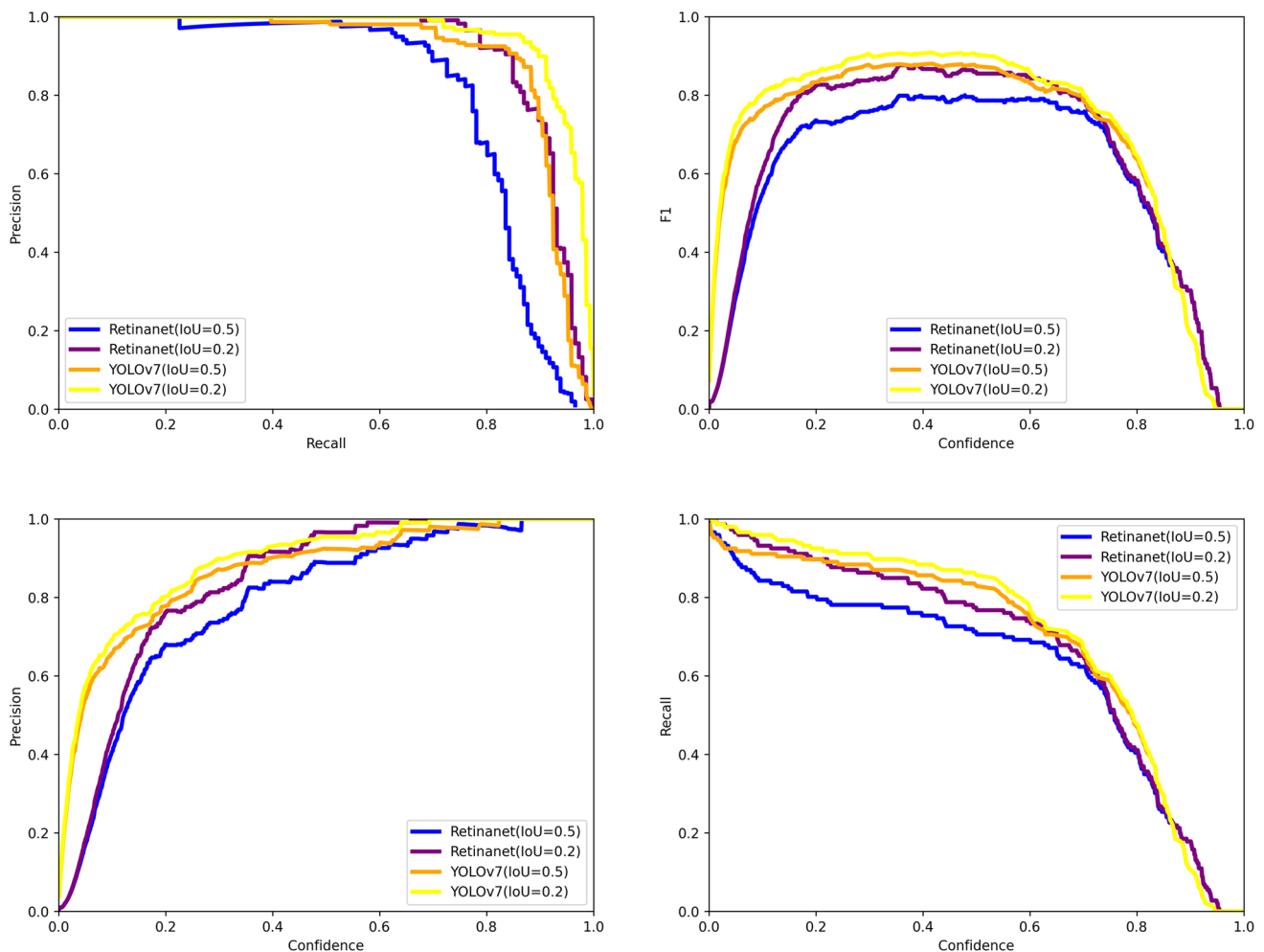


Fig. 6 Evaluation results showing the precision–recall curve and how precision, recall, and the F1 metric change as a function of each model's confidence (see text for details)

Table 2 Results for YOLOv7 when trained from scratch with or without mosaic augmentation

Mosaic	AP
✓	0.905
	0.721

Table 3 Results for RetinaNet with two different backbones when trained from scratch with and without mosaic augmentation, and with and without using the DIoU loss

Backbone	Mosaic	DIoU	AP
ResNet-18-FPN			0.621
ResNet-18-FPN	✓	✓	0.708
ResNet-34-FPN			0.654
ResNet-34-FPN	✓	✓	0.681

a state-of-the-art model and a popular baseline as RetinaNet can be significantly reduced.

Table 4 Influence of focal loss to mitigate the class imbalance problem

Focal loss	AP
✓	0.873
	0.918

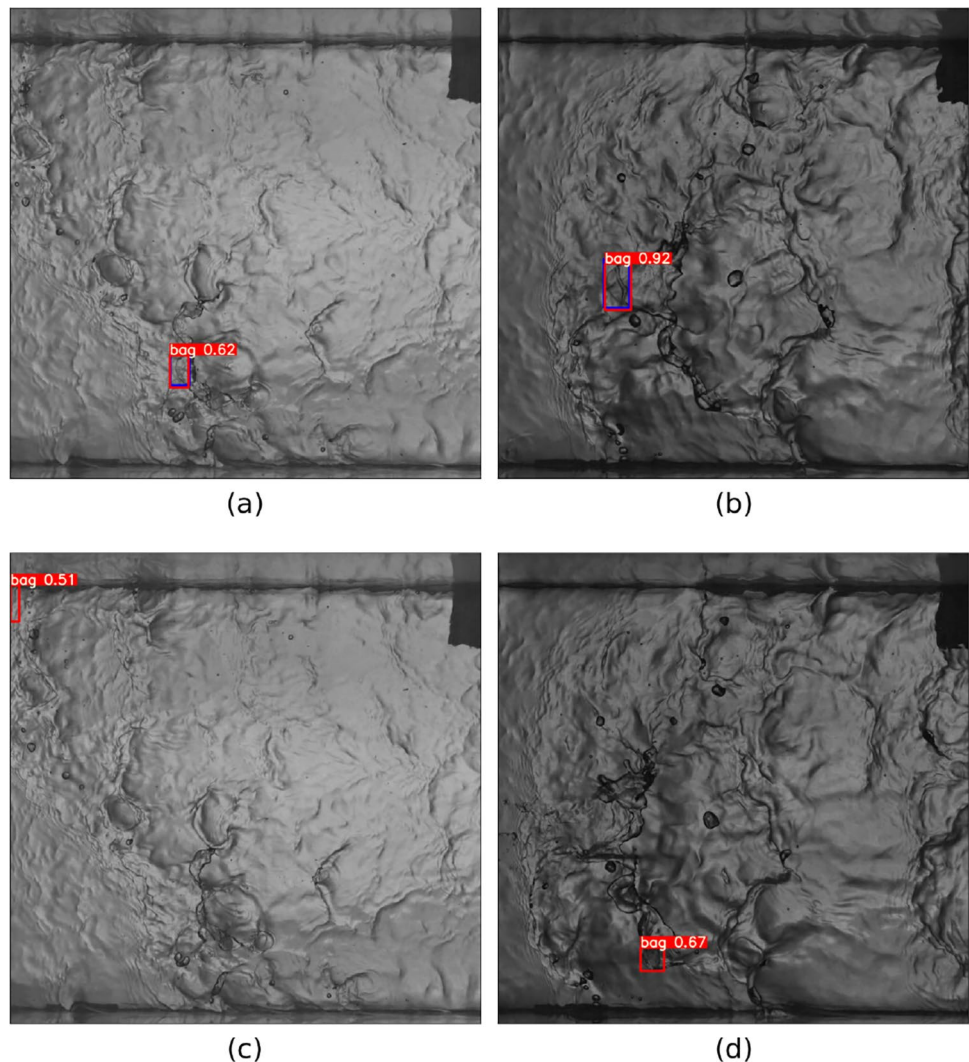
Table 5 Influence of re-parameterization on inference speed

Re-parameterization	Inference speed
	10.2 ms
✓	9.2 ms

4.4.2 Learning from scratch

Although the data set used in this study is very small (about 800 images), by fine-tuning on our data set the pre-trained versions of YOLOv7 and RetinaNet we were able to achieve an Average Precision of more than 0.9 on the test data. To

Fig. 7 Example of YOLOv7 detection. Predictions are shown in red boxes and correct labels in blue boxes. **a, b** Examples of successful detection; **c** An example of false positives at the edge of the image; **d** An example of false detection of an area with similar texture, but not a bag-breakup



investigate which training strategies and components of the models are more effective for a small-sized data set, as the one used in our task, we performed additional experiments by training from scratch the two models under various conditions and evaluated the resulting change in accuracy.

Table 2 shows the experimental results obtained when training YOLOv7 from scratch. The training was performed for 3000 epochs and converged in about 1000–2000 epochs. The table shows that YOLOv7 is able to achieve a very high AP of over 0.9 even without pre-training on a much larger data set. However, when the mosaic augmentation [2] used for training YOLOv7 was removed, the AP decreased by about 0.2 points. This indicates that the mosaic augmentation is very effective for training on small-sized data in this domain.

Next, Table 3 shows the experimental results for learning from scratch when the RetinaNet model was used with two different backbones (ResNet-18-FPN and ResNet-34-FPN). In comparison with the fine-tuning case, now the learning rate was increased 10 times to 1×10^{-4} , and the model was trained for 100 epochs. For both backbones training converged in about 50 epochs. For both ResNet-18-FPN and ResNet-34-FPN the obtained APs were quite low at 0.621 and 0.654, respectively, which is more than 0.2 points less than the accuracy obtained with fine-tuning of the pre-trained model. This indicates that fine-tuning is crucial and the pre-trained weights provide good initial values for the model, even for the present data set, which is from a completely different domain than the natural scene images used for the pre-training.

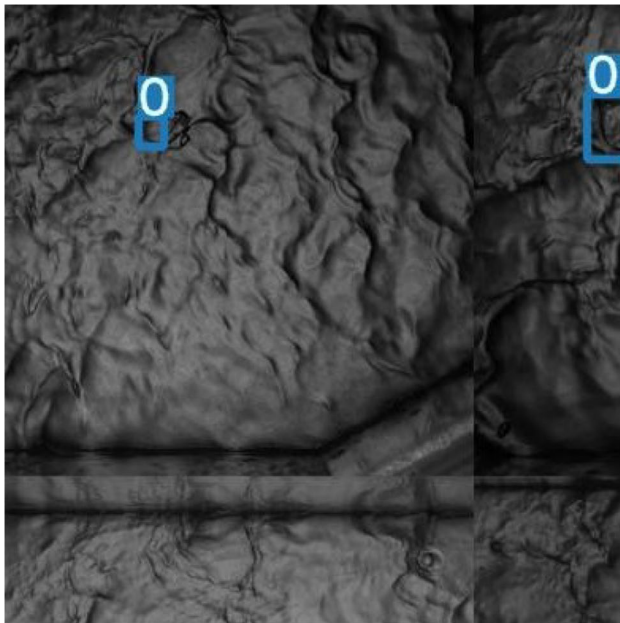


Fig. 8 False positives due to mosaic augmentation. If a box protrudes across the edge of the image, the remaining part inside the image can be wrongly used as a positive target signal

Additionally, we performed further experiments by adding mosaic augmentation, which was important in the case of YOLOv7, and also adding the DIOU loss [22] to the loss function. This improved the accuracy by 8.7% and 2.7% for ResNet-18-FPN and ResNet-34-FPN, respectively, but the accuracy was still lower than that of YOLOv7 without mosaic augmentation.

4.4.3 Influence of the loss function on the class imbalance problem

Since the data set used for our task has a more serious class imbalance between foreground and background than is typical for natural scene images, we measured the change in accuracy resulting when training with and without using the focal loss (a typical countermeasure for mitigating the class imbalance problem). Table 4 shows that the use of focal loss improves the AP by 4.5%, indicating that it is very effective for the data used in our task.

4.4.4 Inference speed

To measure the effect of re-parameterization used in YOLOv7, we compared the inference speed with and without re-parameterization. We used an Intel Xeon W-2223 CPU and an NVIDIA GeForce RTX 3090 GPU to test the inference speed. The image size for inference was 800×800 pixels. The measurement results are shown in Table 5 and indicate that YOLOv7 is capable of very fast inference without re-parameterization, but with re-parameterization it is 1.1 times faster, and real-time processing is possible if a GPU is available.

4.4.5 Qualitative assessment of bag-breakup detection results

Several representative examples of bag-breakup detection when using the YOLOv7 model are shown in Fig. 7. Predictions are shown in red bounding boxes and ground truth annotations in blue boxes. The upper panel of Fig. 7 shows examples of successfully detected bag-breakups, while in the lower row are shown examples of false positives (areas that do not represent a real bag-breakup).

The most common examples of false positives are as those shown in (c), where a small area at the edge of the image is detected incorrectly, and (d), where the texture of the image is very similar to that of a bag-breakup. The former is considered to be due to the use of mosaic augmentation during training. When mosaic augmentation is used, a part of the bounding box may protrude from the image as shown in Fig. 8. In such a case, the remaining part of the bounding box is used as a target signal, which is likely to cause false positives at the edge of the image. Cases like this

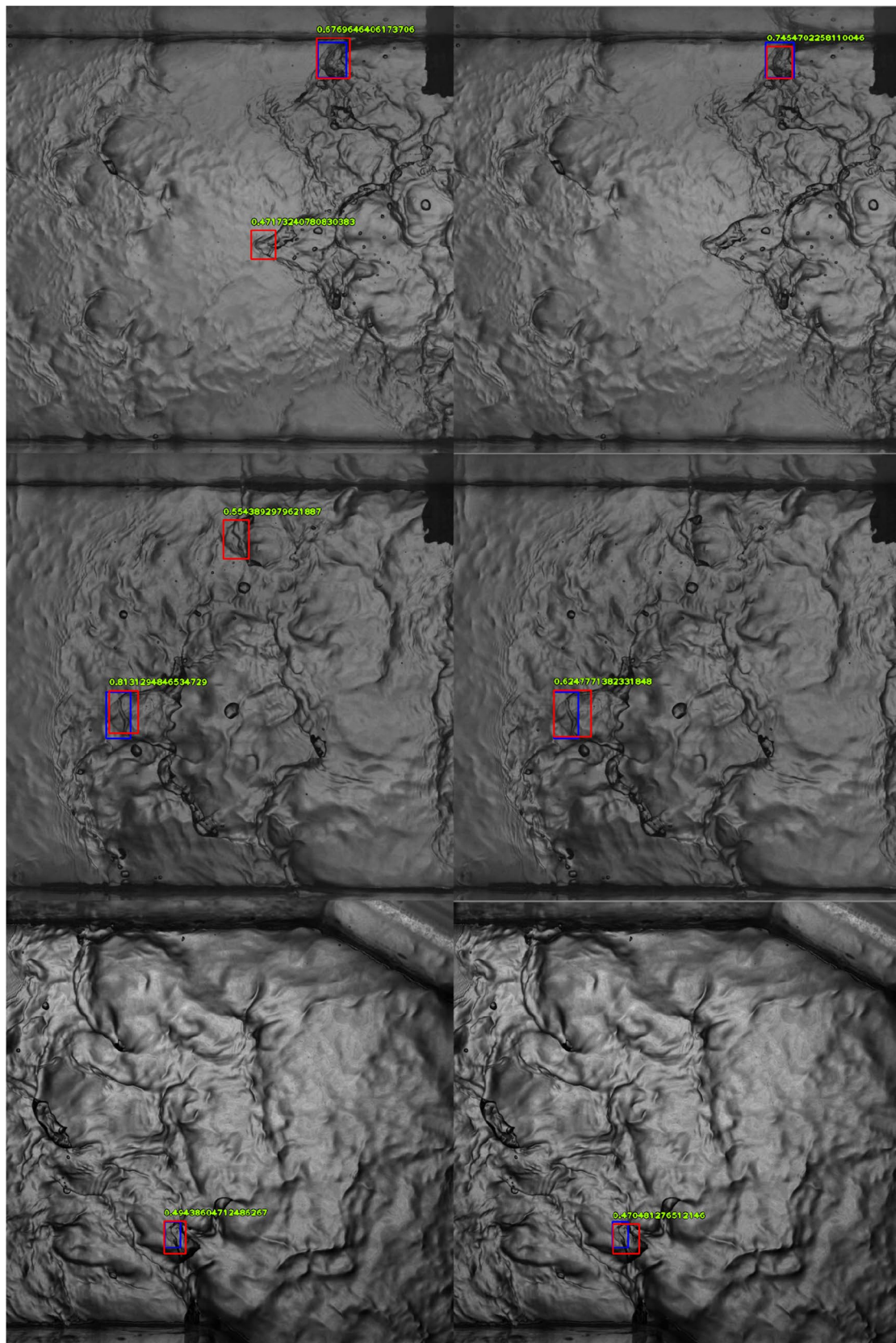


Fig. 9 Several bag-breakup detection examples obtained by RetinaNet when using it as a conventional object detector (left column) and using the proposed method from Sect. 3.4 with $N = 5, i = 4$ (right column). Predictions are shown in red boxes and ground truth in blue boxes. The number above the boxes show model confidence and only detections with confidence above 0.4 are shown. The images in the top two rows show examples where RetinaNet used as a conventional object detector mistakenly detects non-existing bags, while these are successfully avoided when using the method from Sect. 3.4. See main text for further details

can be easily rejected by eliminating small bounding boxes near the edge of the image during post-processing.

The latter example (in Fig. 7d) is difficult even for experts to judge just by looking at this image. False positives of this type can be handled by checking the information from the surrounding image frames, since difficult cases like this one are easier to detect if the sequence of events from the occurrence of a bag-breakup to its disappearance are also considered. In the following subsection, we show how the new method proposed in Sect. 3.4 and inspired by this observation can be used to handle such difficult cases.

4.4.6 Quantitative and qualitative assessment of the detection results obtained by the method from Sect. 3.4

Finally, in this subsection we report the results obtained when using the new method introduced in Sect. 3.4. Table 6 shows the average precision (AP) obtained by RetinaNet for different values of N (number of frames to stack together) and i (length of the intervals before and after the T -th frame, the frame at which detection is performed, see Fig. 4). As can be seen from Table 6, when using our method RetinaNet is able to outperform the conventional YOLOv7 by 1.5%, obtaining for $N = 5, i = 4$ the best AP (0.97) over all conducted experiments. Table 7 shows the Average Precision (AP) obtained by YOLOv7 for different values of N and i . The result show that YOLOv7 also can improve its detection accuracy by using the proposed method, although in this case the improvement is not so big as in the case of RetinaNet (best AP of 0.961 achieved for $N = 5, i = 3$, which is an 0.6% improvement over the best AP obtained when YOLOv7 is used in the conventional way).

Figure 9 shows several bag-breakup detection examples obtained by RetinaNet when using it as a conventional object detector (left column) and when using the proposed method from Sect. 3.4 with $N = 5, i = 4$ (right column). In the figure predictions are shown in red boxes and ground truth in blue boxes. The numbers above the boxes show model confidence and only detections with confidence above 0.4 are shown. The images in the top row and the second row in Fig. 9 show examples where RetinaNet used as a conventional object detector mistakenly detects a bag-breakup

Table 6 Average precision (AP) obtained by RetinaNet when using information from neighboring frames for different values of N and i

i	$N - 1$	N	$N + 1$	$N + 2$
N				
3	0.954	0.947	0.942	0.929
5	0.970	0.895	0.924	0.873
7	0.904	0.920	0.846	0.830

Table 7 Average precision (AP) obtained by YOLOv7 when using information from neighboring frames for different values of N and i

i	$N - 1$	N	$N + 1$	$N + 2$
N				
3	0.941	0.961	0.958	0.956
5	0.927	0.894	0.942	0.910
7	0.854	0.909	0.930	0.871

in areas of the images which closely resemble a bag. Without utilizing additional information from the neighboring frames, i.e., without looking at several consecutive frames, it is difficult even for human experts to determine whether these areas represent bags or not. However, using the proposed method from Sect. 3.4 (as the corresponding results in the right column of Fig. 9 show) it is possible to eliminate these false positives. In the lack of such ambiguous regions in the images, as is the case in the images in the bottom row of Fig. 9, both methods perform equally well and there are no false positive detections.

5 Conclusion

In this paper, we have shown that the task of automatic detection of bag-breakup in images of droplet scattering can be accomplished with very good accuracy and in real-time using state-of-the-art object detection methods, even though the available training images were less than 600 and the image texture for this fluid phenomenon has very different characteristics from those of natural scene images. We have addressed several important challenges which practitioners in scientific or industrial research settings might face in their special domain applications and have shown several strategies which can be used successfully to improve accuracy in such settings, at the same time providing both quantitative and qualitative evaluations of their effects.

Additionally, we have proposed a new method, which by utilizing information from several consecutive frames further improves accuracy of detection by eliminating false positives which might occur in regions which closely resemble the visual structure of bags and are difficult to discriminate from a single image. We hope that the practical insights gained in this study can be of use to other researchers and

practitioners when targeting applications where the images differ greatly from natural images.

Acknowledgements This research was supported by Mazda Motor Corporation. The droplet dispersal related experiments from which the images were generated were carried out at the Mechanical Power and Motor Systems Laboratory, Hiroshima University. B.R. is supported in part by a JSPS KAKENHI Grant Number JP23K11170.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alessio, H., Bassett, D., Bopp, M., Parr, B., Patch, G., Rankin, J., Rojas-Rueda, D., Roti, M., Wojcik, J.: Climate change, air pollution, and physical inactivity: Is active transportation part of the solution? *Med. Sci. Sports Exerc.* **53**(6), 1170–1178 (2021)
- Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
- Cherapanamjeri, J., Rao, B.N.K.: Neural networks based object detection techniques in computer vision. In: 4th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1092–1099 (2022)
- Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)
- Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep* **2007**, 1–45 (2012)
- Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R.: A survey of deep learning-based object detection. *IEEE Access* **7**, 128837–128868 (2019)
- Kaur, R., Singh, S.: A comprehensive review of object detection with deep learning. *Digital Signal Process.* **132**, 103812 (2023)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
- Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125 (2017)
- Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740–755. Springer (2014)
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. *Int. J. Comput. Vision* **128**(2), 261–318 (2020)
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, pp. 21–37. Springer (2016)
- Nakada, S., Akiyama, K., Ma, J., Nishida, K., Yamamoto, R., Nakashima, A., Nakamura, K., Marui, K., Nishimura, M., Yokohata, H., Ogata, Y.: Study of breakup on water film sheared by steady and pulsatile air flow in a horizontal rectangular duct. The 31st International Symposium on Transport Phenomena, 13–16 October 2020, Honolulu, USA (2020)
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016)
- Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **28**, 5 (2015)
- Saponara, S., Elhanashi, A.: Impact of image resizing on deep learning detectors for training time and model performance. In: Applications in Electronics Pervading Industry, Environment and Society, pp. 10–17. Springer International Publishing (2022)
- Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10778–10787 (2020)
- Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint [arXiv:2207.02696](https://arxiv.org/abs/2207.02696) (2022)
- Wang, C.Y., Liao, H.Y.M., Yeh, I.H.: Designing network design strategies through gradient path analysis. arXiv preprint [arXiv:2211.04800](https://arxiv.org/abs/2211.04800) (2022)
- Xiao, X., Gao, M.: Overview of climate change, air pollution, and human health. In: M. Gao, Z. Wang, G. Carmichael (eds.) *Air Pollution, Climate, and Health*, pp. 3–12. Elsevier (2021)
- Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34–7, pp. 12993–13000 (2020)
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint [arXiv:2010.04159](https://arxiv.org/abs/2010.04159) (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.