Supplemental Material for ComplexVAD: Detecting Interaction Anomalies in Video

Furkan Mumcu*
University of South Florida
furkan@usf.edu

Michael J. Jones
Mitsubishi Electric Research Labs (MERL)
mjones@merl.com

Yasin Yilmaz
Univ. of South Florida
yasiny@usf.edu

Anoop Cherian
Mitsubishi Electric Research Labs (MERL)
cherian@merl.com

1. More details on ComplexVAD

1.1. Dataset Details

In this section, we provide additional details about the dataset.

Motivation The ComplexVAD dataset was created to encourage new solutions to the video anomaly detection problem, and in particular, to encourage methods that can handle anomalous interactions among people and objects which often occur in real-world scenarios. The collection and labeling of the dataset was done as a collaboration by researchers at the University of South Florida and Mitsubishi Electric Research Laboratories and funded by the University of South Florida and Mitsubishi Electric Research Laboratories.

Composition The dataset is comprised of three directories. The "Train" directory contains 104 MPEG videos of a single scene taken in a public space on the campus of the University of South Florida (USF). The videos in the Train directory define normal activity for this scene. The scene shows a two-lane street with a pedestrian crosswalk going across it as well as sidewalks on either side of the street. Car parking lots are also visible in the background. The "Test" directory contains 113 MPEG videos of the same scene on the USF campus. Videos in the "Test" directory contain one or more anomalous activities such as a person leaving behind a package, a cyclist colliding with a pedestrian or a person sitting on the hood of a car. The "annotations" directory contains 113 JSON files (one for each test video) with ground truth annotations for all anomalies in each test video. The format of an annotation file is as follows:

```
    {
"total_frame": ...,
"annotations": [
        {
        "track_id": ...,
        "frame_id": ...,
        "bbox": ...,
        "object_type": ...
        },
        {
        "track_id": ...,
        "frame_id": ...,
        "bbox": ...,
        "object_type": ...
        },
        ...]
}
```

where total_frame represents the total number of frames in the video. The annotations field contains the list of each annotated object in every frame with the following properties:

track_id: unique id for the object
frame_id: frame number of the object
bbox: bounding box of the object in the format of [x1, y1, x2, y2] where (x1, y1) is the coordinate of top-left and (x2, y2) is the coordinate of top-right for the bounding box
object_type: type of the object i.e., person, skateboard.

Note that a unique track_id represents the same object through different frames. If a particular object is present in consecutive frames, the corresponding annotations will have the same track_id with different frame_id and bbox values.

Videos were collected at various times during the day and on each day of the week. Videos vary in duration with most being about 12 minutes long. The total duration of all training and testing videos is a little over 34 hours. Each frame has a resolution of 1920 pixels wide by 1080 pixels high.

The videos in the Train directory should be used to learn a model of normal activity for the scene. Videos in the Test directory should be used for trying to detect anoma-

[*] Furkan Mumcu did part of this work as an intern at MERL.

lous activity (activity that does not occur in any training video). The annotations are used for evaluating the accuracy of anomaly detection using the region-based detection criterion [5], track-based detection criterion [5] or frame-level criterion [4].

Collection Process All videos were collected using a Canon EOS Rebel T6 video camera set on a tripod on the USF campus. Videos are stored as MPEG files using an MPEG-H Part 2/HEVC (H.265) (hev1) codec. Frame resolution is 1920x1080 pixels and videos are recorded at 30 frames/second. Videos were collected over many different days over a 5 month period in 2023. On each day that video was collected, the camera was positioned in approximately the same way so that approximately the same area was in view for every video. For nominal videos in the Train directory, the camera simply recorded naturally occurring activities in the scene. For videos in the Test directory, some videos were acquired from naturally occurring activity that happened to capture unusual events while others were acquired by actors who purposely created anomalous interactions.

The Institutional Review Board at USF was consulted about the collection of video in a public space and concluded that because the "project does not include interacting with the individuals in the recordings to collect information, then it does not meet the definition of Human Subjects Research and does not require submission of an application for the IRB's review."

Preprocessing/cleaning/labeling In order to preserve the privacy of people captured in the videos, a face detector [2] was run on every frame and any detected faces were blurred with a Gaussian kernel.

The annotations for all anomalies in the Test videos consist of bounding boxes around each person/object involved in the anomalous activity as detailed above. The annotations were manually created using the Computer Vision Annotation Tool (CVAT) (https://www.cvat.ai).

Distribution The ComplexVAD dataset can be freely downloaded from:

https://www.merl.com/research/downloads/ComplexVAD. It is distributed under the CC-BY-SA-4.0 license.

Maintenance ComplexVAD is maintained by Mike Jones at MERL who can be contacted regarding any questions about the dataset.

1.2. Anomaly Types

The ComplexVAD dataset includes many different types of anomalies, many of which involve interactions among two objects or actors. Figure 1 demonstrates the numbers of each anomaly type represented in the ComplexVAD for the following list of anomaly types represented in the dataset:

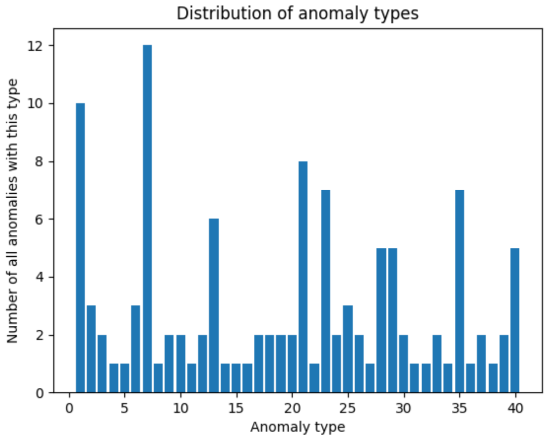


Figure 1. Numbers of each anomaly type represented in the ComplexVAD. The numbers along the x-axis are the anomaly type indices listed in the paper. The top three most common anomaly types are: 7- Person falling from a skateboard, 1 - Person leaving an object on the ground, and 21 - Skateboarder uses the main road.

1. Person leaving an object on the ground
2. Person blocking a car
3. Car hitting a person
4. Bicycle hitting a person
5. Person trying to break into a car
6. Person sitting on a car
7. Person falling from a skateboard
8. Piggyback
9. Dog not on leash
10. Two people fighting
11. Person pushing someone to the road
12. Runner colliding with another person
13. Car breaking hard to stop for pedestrian
14. After stopping for pedestrian car unexpectedly moves, which makes the pedestrian run
15. Pedestrian preparing to cross the street has to stop because car does not stop
16. Imitating vandalizing a car (e.g., with a long stick or baseball bat)
17. Person hitting a tree with baseball bat
18. Person nailing something to a tree
19. Multiple people suddenly running scattered around

20. Skateboard moves on its own without a user

21. Skateboarder uses the main road

22. Person hits someone with a bat, takes his wallet, then runs

23. People play with a ball in the middle of the street

24. Students kick soccer ball across the street

25. Two men carry bike

26. Two bikers bump each other

27. Two skateboarders bump each other

28. Two people ride one scooter

29. Scooters, bikes left alone

30. Person leaves an object on top of car

31. Person walking with unusual path

32. Man tries to climb a pole

33. A golf cart with a trailer stops and waits

34. Woman pushing a trolley

35. Person pushes a skateboard with his feet while skateboard has bag on it

36. Person carries another person with a trolley

37. Person ties shoelace in the middle of the street

38. Person falling while running/walking

39. Biker going on a non-straight path (e.g., taking a u-turn)

40. Skateboarder going on a non-straight path (e.g. u-turn)

1.3. Distribution of objects in ComplexVAD dataset

To give some more insight into the contents of the ComplexVAD dataset, Figure 2 shows bar graphs of the number of detections for the top 8 object classes detected in the Train and Test videos. The Detectron2 [7] object detector which was trained on the 80 classes from MS-COCO [3] was used to detect objects in each frame of the Train and Test videos. There were a total of 38,754,900 objects detected in the Train videos and 28,847,159 objects detected in the Test videos. Cars are the most common object detected due to the parking lot in the background of the scene and people are the second most common object class.

2. Further details on model building

Normalization Constants

The five attribute distances in Equation (8) need to have similar scales so that one does not dominate the others. To insure this, each attribute distance is normalized by subtracting the mean and dividing by the standard deviation. We use pairs of nodes computed from the nominal video of a dataset to compute each attribute's distance distribution for that dataset. The resulting normalized distances are less than 0 if two nodes are very similar (raw attribute distance less than the mean), and greater than 1 if two nodes are significantly different (raw attribute distance greater than the mean plus standard deviation).

Selecting exemplars across all nominal videos

The model building process described so far selects sets of exemplars (for isolated nodes and node pairs) for a single nominal video. Because most datasets, including ComplexVAD, include multiple nominal videos, we need a way of selecting exemplars across all nominal videos. To do this, we simply take the union of all the exemplar sets selected for each nominal video (again, independently for isolated nodes and node pairs). Then we run exemplar selection again over the union set. This effectively removes similar exemplars in the union set and leaves a final set of exemplars that cover the variety of exemplars found in all nominal videos. The final result is a set of isolated node exemplars denoted $\mathcal{E}_{iso}$ and a separate set of node pair exemplars denoted $\mathcal{E}_{pair}$ across all nominal videos.

3. Visualizations of results

We have included visualizations of the anomaly detections for our new method as well as the EVAL [6] and MemAE [1] methods on subsequences from 6 different test videos from the ComplexVAD dataset. Each subsequence contains an anomalous event. [1]

The result videos for our method (filenames beginning with "Ours") and for the MemAE method (filenames beginning with "MemAE") show green bounding boxes around annotated ground truth anomalies and red bounding boxes around detected anomalies. The MemAE result videos are much lower resolution and are grayscale because this is the input to the MemAE algorithm. The result videos for the EVAL method show regions detected as anomalous shaded in red. Ground truth annotations are not visualized in the EVAL result videos.

We will discuss each result video individually below, but we first make some general comments. The result videos

[1] https://merl.com/research/highlights/ComplexVAD/2_supp.zip

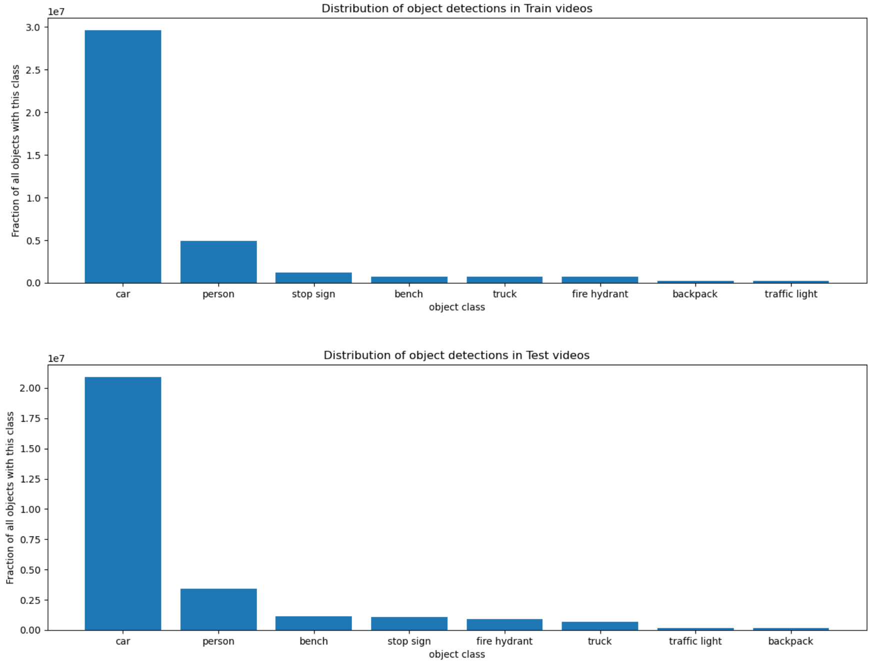


Figure 2. Numbers of each object class detected in the Train (Top) and Test (Bottom) videos for the top eight classes.

show that our method generally does a good job in detecting anomalous activity and has relatively few false positive detections. In some cases, our method detects anomalous activity that is not marked in the ground truth annotations but can reasonably be regarded as anomalous. For example, a person loitering around the cross-walk when there are no cars coming. This did not occur in the nominal videos, but was not marked as anomalous in the ground truth annotations. Manually annotating anomalies is difficult due to many ambiguous cases. For the EVAL method, it detects many of the anomalies, but also has many more false positives than our method. Furthermore, its localization of anomalies is much looser than ours due to EVAL's use of a grid of fixed-sized regions instead of the object detections that we use. For the MemAE method, it does a poor job of localizing anomalies and also has very many false positives especially in the tree branches for which there is a lot of movement due to wind.

In the following, we discuss each result video individually.

video 4344: This video shows a person crossing the street at the cross-walk and then suddenly kneeling down in the middle of the street. The person then gets back up and continues walking. Our method does a good job of detecting this anomalous activity both temporally and spatially with no false positives. The EVAL method also detects well with no false positives although its detections are much looser around the person. The MemAE method fails to detect the anomaly and has many false positive detections in the swaying tree branches.

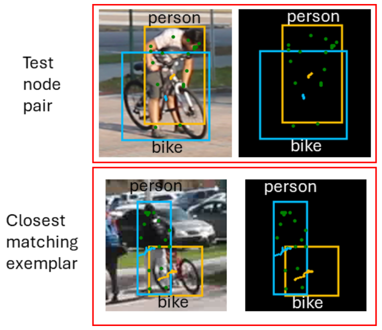
video 4371: This video shows a person on a bike and a person on a scooter (slowly) bump into each other in the middle of the cross-walk and then go around each other to continue moving across the street. Our method detects both pairs of objects (person-bike and person-scooter) for much of the anomalous interaction. It only has a few, small false positive detections on people walking at the right of the frame near the end. The EVAL method also detects the anomalous activity, but has quite a few false positive detections in the trees and other areas as well as continuing to detect the person and bike while they are moving normally after the anomalous interaction. The MemAE method fails to detect the anomalous event and has many small false positives especially in the swaying tree branches.

video 4376: This video shows a person loitering on the sidewalk in front of the cross-walk and then a person riding a bike nearly runs into him. The person moves out of the way and the biker continues across the street. Our method detects a good proportion of the anomalous activity as anomalous. It also detects the person loitering as anomalous. Even though this is not marked as a ground truth anomaly, it can be considered anomalous because it does not occur in the nominal videos. Our method has a few small false positives on a car in the background. The EVAL method also does a good job of detecting the anomalous interaction and also detects some instances of the person loitering, but it has many more false positives than our method. The MemAE method again fails to detect the anomaly and has many false positives.

video 4379: This video shows a biker riding across the street in the cross-walk, but then unexpectedly stopping in the middle of the cross-walk before continuing across the street. Our method does a good job of correctly detecting the stopped biker with no false positives. The EVAL method also detects the anomaly well, but has a few false positives. The MemAE method has a few very small detections on the stopped biker but does a poor job of spatially localizing this anomaly. It also continues to have many false positive detections.

video 4383: This video shows a person walking his bike across the street and then stopping and parking the bike on the sidewalk and then walking away from the bike. The ground truth annotation marks the person stopping and parking his bike as anomalous as well as marking the left-behind bike as anomalous. Our method detects some of the instances of the person stopping and parking his bike as anomalous and also detects the left-behind bike as anomalous. There are a couple of short-lived false positive detections. The EVAL method fails to detect any of the anomalous activity (parking the bike on the sidewalk or leaving the bike behind) and has a larger number of false positives. The MemAE method does not detect the anomalous activity and has many false positives.

video 4398: This video shows a person loitering on the sidewalk with a soccer ball. Then a skateboarder comes and runs into the soccer ball followed by the skateboarder, the soccer ball and the person all crossing the street. Here once again, what to annotate as anomalous is ambiguous. Only the skateboarder running into the soccer ball is marked as anomalous. However, the person loitering with the soccer ball and the skateboarder and soccer ball crossing the street near each other could also be considered anomalous. Our method correctly detects much of the annotated anomaly but also detects the person and the soccer ball that are stationary at the beginning of the video as anomalous. It also detects the skateboarder and soccer ball traveling together across the street as anomalous. The EVAL method fails to detect most of the skateboarder running into the soccer ball. It does detect the person loitering at the beginning as well as some of the person and soccer ball crossing the street which is arguably anomalous. EVAL also has a few more false positives than our method. The MemAE method once again does not detect the anomalous activity and continues to have many small false positives all around the image and especially in the swaying trees.



Attribute distances (for correct correspondence between the pairs of test and exemplar objects):

Location distances = -.08, .20
Class distances = 0, 0
Trajectory distances = .57, .71
Size distances = .17, .35
Pose distances = .03, 0

Overall distance = 0.71

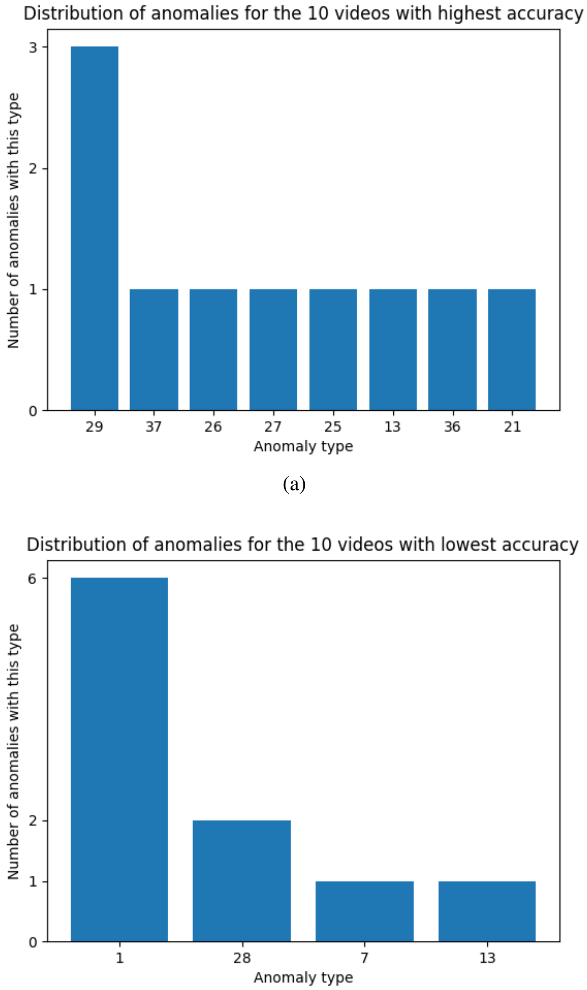
Figure 3. Visualization of a test node pair with four of its attributes: class ID, trajectory, size, pose (the location within the frame is not visualized here).

4. Visualizations of object attributes and closest matching exemplar

Figure 3 shows in the top, left a pair of interacting objects (person and bike) from a test video that are detected by our method and linked due to proximity. This test node pair represents an anomalous person and bike that have stopped in the middle of the road. Below the test node pair is a visualization of the closest matching exemplar node pair. The closest matching exemplar pair is also a person and bike but the person is walking the bike toward the left of the frame. For each object, the size is indicated by the bounding box and the class ID is written above or below the box. The trajectory for each object is visualized by a sequence of 30 dots (one for each of the 30 frames that it is tracked) of the same color as the object's bounding box and starting from the middle of the box. For these test objects which are barely moving, the trajectory is very short. The 17 coordinates that comprise the pose of a person are shown as green dots. The same set of attributes are shown on the left of each visualization overlayed on the original frame and then again on the right over a black background so that they are more easily seen. The figure also shows the attribute distances between the test node pair and the closest matching exemplar pair for the correct correspondence between objects in the pairs. From this we can see that the trajectory distance (0.71) is the largest and is assigned as the anomaly score according to Equation 8 in the main paper. This distance is above the anomaly threshold of 0.5 which results in the test person and bike being detected as anomalous. This means that there was no pair of person and bike found in the nominal training videos with a similar trajectory (stationary). This is because all people and bike pairs in the nominal videos were moving across the road and not stopped in the middle of the road. We can easily use this information to provide a simple explanation of why this pair of person and bike was detected as anomalous.

5. More insights about our method's performance on ComplexVAD

Figure 4 shows the distribution of anomalies for the 10 videos with highest (a) and lowest (b) accuracy for our method. According to the results, the most common anomaly



(a)

(b)

Figure 4. Distribution of anomalies for the 10 videos with highest (a) and lowest (b) accuracy for the Scene-Graph method. The most common anomaly type in the highest-accuracy videos was 29 - Scooters, bikes left alone. The most common anomaly type in the lowest-accuracy videos was 1 - Person leaving an object on the ground.

type in the highest-accuracy videos was 29 - Scooters, bikes left alone while the most common anomaly type in the lowest-accuracy videos was 1 - Person leaving an object on the ground.

Crowded scenes is one of the challenging aspect of ComplexVAD dataset and may cause difficulties for object based methods which detect and track objects. Figure 5 shows false positives raised by our method.

In addition to interaction based complex anomalies, ComplexVAD also includes non-interaction-based simple anomalies, such as the example shown in Figure 6. In this specific example, the skateboarder goes on the road, which is a simple non-interaction anomaly, similar to the location-based anomalies commonly found in the existing datasets. The Scene Graph method addresses simple anomalies by analyzing single objects. Equation (10) is the distance function, which is primarily designed to detect simple anomalies.

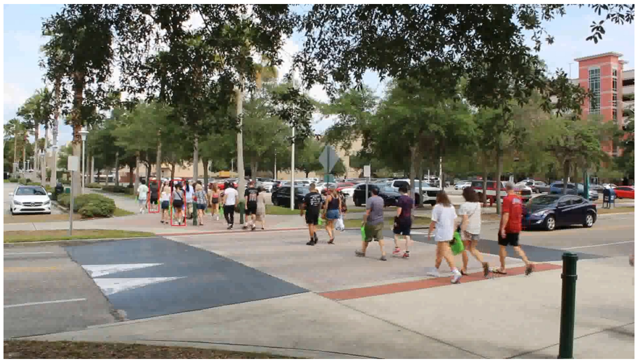


Figure 5. Crowded scene example. Red bounding boxes show the false positives our method raises momentarily.

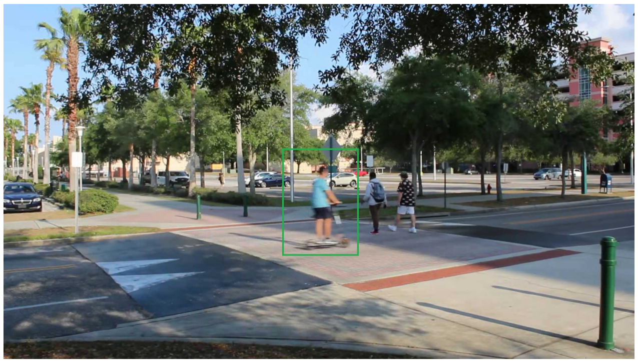


Figure 6. Simple anomaly example. Skateboarder goes on the road. Similar location-based simple anomalies are commonly found in the existing datasets.

References

[1] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[2] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5060–5069, 2019. 2

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference,*

Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 3

[4] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1975–1981. IEEE, 2010. 2

[5] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020. 2

[6] Ashish Singh, Michael J. Jones, and Erik Learned-Miller. Eval: Explainable video anomaly localization. In *CVPR*, 2023. 3

[7] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 3