ECCV 2022 Seasons in Drift Challenge

Fact sheet

This is the fact sheet's template for the ECCV 2022 Seasons in Drift Challenge. Please fill out the following sections carefully in a scientific writing style. Then, send the compressed project (in `.zip` format), i.e., the generated PDF, `.tex`, `.bib` and any additional files, following the schedule and instructions ("Wining solutions (post-challenge)", Fact Sheets) provided in the Challenge webpage.

I. TEAM DETAILS

- Team leader name: Xiaoqiang Lu
- Username on Codalab: GroundTruth
- Team leader affiliation: Xidian University
- Team leader email: xqlu@stu.xidian.edu.cn
- Name of other team members (and affiliation): Tong Gou (Xi'an University of Technology), Zhongjian Huang (Xidian University), Yuting Yang (Xidian University), Xu Liu (Xidian University), LingLing Li (Xidian University), Fang Liu (Xidian University), Licheng Jiao (Xidian University)
- Team website URL (if any):
- Competition track (mark with X one single option):
  - (X) Track 1: Detection at **day** level.
  - ( ) Track 2: Detection at **week** level.
  - ( ) Track 3: Detection at **month** level.

II. CONTRIBUTION DETAILS

A. Title of the contribution

For video object detection, one of the simplest methods is to use a still object detector to train every frame from the video. However, due to a large amount of redundant information between frames of video, and the drawbacks of excessive data, this undoubtedly leads to computationally expensive training of still object detectors and low detection accuracy. In addition, there is temporal and contextual information in the video data compared to the still images. Obviously, using this information will help the detector detect the object better and faster. To this end, we first explore where the limit of the still object detector is. Based on Scaled-YOLOv4 [1], we first perform sparse sampling at the input. According to different sampling methods, the best sampling settings are obtained through a lot of experiments. Secondly, we use Mosaic [2] data augmentation to better improve the detector's recognition ability and robustness to small objects. In inference, we first adopt Model Soups [3] for model integration to obtain a more accurate and robust model. In addition, simply adding a test time augmentation of horizontal flipping improves detection performance even further. Finally, applying Seq-NMS [4] post-processing to the detection results of the still object detector, we achieve the best performance in the ECCV'22 ChaLearn Seasons in Drift Challenge Track 1 [5].

B. Representative image / workflow diagram of the method

Our framework is shown in Figure 1.

C. Detailed method description

1) ***Improving Scaled-Yolov4 for Video Object Detection with Sparse sampling and Seq-NMS****:* We first perform sparse sampling at the input. Common sampling methods include average sampling and random sampling, with sampling rates set to 0.5, 0.2 and 0.1 respectively. In addition, inspired by active learning, we also carried out active sampling to select the key frames of the video. To be specific, we used the average confidence score of all detection boxes in an image as the uncertainty of the image, then sorted it in ascending order according to the uncertainty, and finally selected the top 0.5, 0.2 and 0.1 images according to the sampling rate for training. As shown in Table I, the average sample with a sampling rate of 0.2 performed best.

Given a video sequence of region proposals and their corresponding class scores, Seq-NMS [4] associates bounding boxes in adjacent frames using a simple overlap criterion. It then selects boxes to maximize a sequence score. Those boxes are then used suppress overlapping boxes in their respective frames and are subsequently rescored in order to boost weaker detections. We apply Seq-NMS to the detection results and improve the performance further.

2) ***Data pre-processing****:* The training dataset of the day-level contains a total of 6360 frames, and 4385 frames are not empty. Thus, we only trained on the frames which have objects. Besides, due to the class imbalance in the dataset, this is reflected in the low number of frames containing bicycle or motorcycle. Therefore, sparse sampling will result in fewer samples of these two categories. To do this, we sampled frames containing bicycle, and the sampled frames and all frames containing motorcycle were used to expand the dataset.

3) ***Implementation details****:* For training, we use Scaled-YOLOv4p6 and Scaled-YOLOv4p7. For Scaled-YOLOv4p6, a single-scale training with an input image size of 1280x1280 and a batch size of 32 are used. For Scaled-YOLOv4p7, a single-scale training with an input image size of 1536x1536 and a batch size of 16 are used. The time for training is

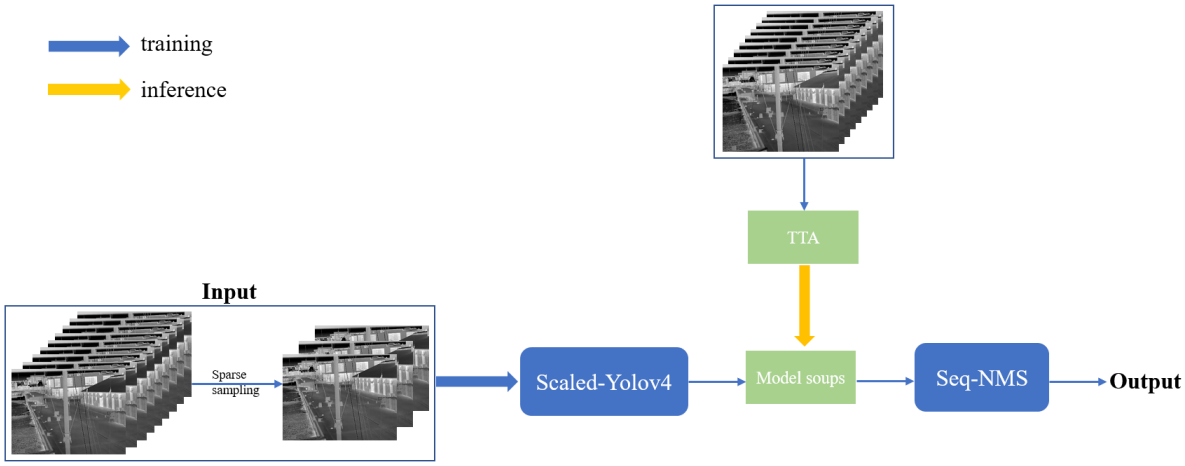

Fig. 1. Our method framework.

TABLE I
THE EXPERIMENTAL RESULTS ON VALID DATASET.

Sample rate	Sample method	mAP
1	-	0.230
0.1/0.2/0.5	average	0.258/**0.263**/0.244
0.1/0.2/0.5	random	0.249/0.251/0.242
0.1/0.2/0.5	active	0.253/0.256/0.242

100 epochs. A SGD optimizer with an initial learning rate of 0.01, and a weight decay of 0.0005 are used.

For inference, the input image size is same as the training setting. Test time augmentation only contains horizontal flipping. And the iou threshold of NMS if 0.5, the confidence threshold is 0.1.

D. Challenge results

Our results in test phase is shown in Table II.

E. Final remarks

Our proposed method is a simple yet effective approach to improving still object detector for video object detection. But we only consider applying temporal and contextual information to the detection results. Thus, we will explore using this information to apply to feature space in the future.

III. ADDITIONAL METHOD DETAILS

Please, reply if your challenge entry considered (or not) the following strategies and provide a brief explanation. For each question, mark with X one single option.

- **For the competition track associated with this fact sheet, you confirm that you have trained your model on the predefined and single:** (X) Day, ( ) Week, ( ) Month - as instructed in the challenge webpage.
- **Did you use any pre-trained model:** (X) Yes, ( ) No. If yes, please detail: We use the model pre-trained on COCO dataset provided by `https://github.com/WongKinYiu/ScaledYOLOv4`.
- **Did you use external data?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you perform any data augmentation?**
  (X) Yes, ( ) No
  If yes, please detail: Mosaic.
- **At the final phase, did you use the provided validation set as part of your training set?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any regularization strategies/terms?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use handcrafted features?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any spatio-temporal feature extraction strategy?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you perform object tracking?**
  ( ) Yes, (X) No
  If yes, please detail:
- **Did you leverage timestamp information?**
  ( ) Yes, (X) No
  If yes, please detail:

TABLE II

RESULTS FROM LEADERBOARD (TEST PHASE) OBTAINED BY THE PROPOSED APPROACH.

Rank position	mAP_w	mAP	Jan	Mar	Abr	May	Jun	Jul	Aug	Sep
1	0.279846	0.2832	0.3048	0.3021	0.3073	0.2674	0.2748	0.2306	0.2829	0.2955

- **Did you use empty frames present in the dataset?**
  ( ) Yes, (X) No
  If yes, please detail:

- **Did you construct any type of prior to condition for visual variety?**
  ( ) Yes, (X) No
  If yes, please detail:

IV. CODE REPOSITORY

Link to a code repository with complete and detailed instructions so that the results obtained on Codalab can be reproduced locally. This includes a list of requirements, pre-trained models, and so on. Note, training code with instructions is also required. This is recommended for all participants and mandatory for winners to claim their prize. **Organizers strongly encourage the use of docker to facilitate reproducibility**.

Code repository: `https://github.com/xiaoqiang-lu/seq-yolo`

REFERENCES

[1] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2021, pp. 13 029–13 038.

[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[3] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," *arXiv preprint arXiv:2203.05482*, 2022.

[4] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[5] I. Nikolov, M. Philipsen, J. Liu, J. Dueholm, A. Johansen, K. Nasrollahi, and T. Moeslund, "Seasons in drift: A long-term thermal imaging dataset for studying concept drift," in *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.